THE BELL SYSTEM TECHNICAL JOURNAL

SEPTEMBER 1978
VOL. 57, NO. 7, PART 2

ISSN0005–8580

THE BELL SYSTEM TECHNICAL JOURNAL

THE BELL SYSTEM TECHNICAL JOURNAL is published monthly, except for the May-June and July-August combined issues, by the American Telephone and Telegraph Company, J. D. deButts, Chairman and Chief Executive Officer; C. L. Brown, President; W. G. Burns, Vice President and Treasurer; F. A. Hutson, Jr., Secretary, Editorial enquiries should be addressed to the Editor, The Bell System Technical Journal, Bell Laboratories, 600 Mountain Ave., Murray Hill, N.J. 07974. Checks for subscriptions should be made payable to The Bell System Technical Journal and should be addressed to Bell Laboratories, Circulation Group, Whippany Road, Whippany, N.J. 07981. Subscriptions $20.00 per year; single copies $2.00 each. Foreign postage $1.00 per year; 15 cents per copy. Printed in U.S.A. Second-class postage paid at New Providence, New Jersey 07974 and additional mailing offices.

THE BELL SYSTEM
TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 57 September 1978 Number 7, Part 2

An Approximation for the Variance of the UPCO Offered Load Estimate

By E. J. MESSERLI

(Manuscript received September 26, 1977)

This paper develops a generalization of some available approximations for the variance of the estimate for offered load to a trunk or server group operating in a blocked-calls-cleared mode, using measurements of usage, offered attempts (peg count), and overflow. The analysis takes into account the peakedness of the offered traffic stream, the level of blocking on the group, the duration of the measurement interval, and switch count errors due to sampling usage. The resulting approximation is quite accurate over a wide range of conditions, is easily computable, and clearly displays the role of the basic factors that control the precision of the estimator. The variance approximation is useful in studies of the relationship between traffic measurement errors and the performance of the provisioning and administration processes.

I. INTRODUCTION

The estimation of loads offered to a trunk group or server group operating in a blocked-calls-cleared mode plays an important role in many network-provisioning processes. The preferred measurement combination for developing such load estimates consists of usage, offered attempts (peg count), and overflow attempts (usually referred to in the Bell System as UPCO measurements). This paper develops a generalization of some available approximations for the variance of the UPCO offered load estimate for a single measurement interval. The analysis considers the peakedness of the offered traffic stream, the level of blocking or call congestion for the group, the duration of the measurement interval, and switch count errors due to the sampling of usage at discrete points in time. The resulting approximation is quite accurate

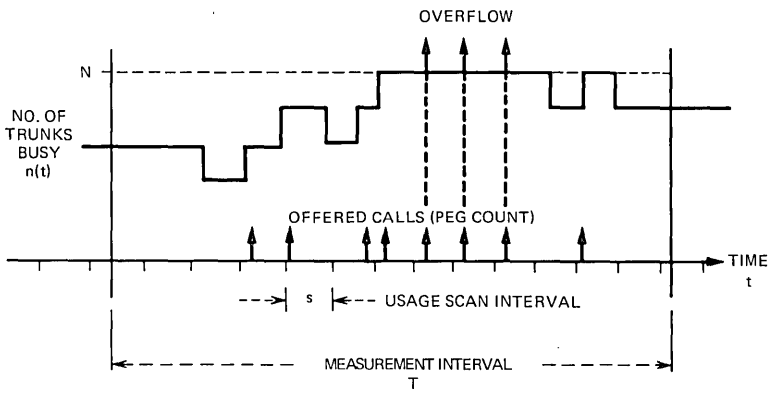
over a wide range of conditions, is easily computable, and clearly displays the role of the basic factors that control the precision of the estimator.

Variance approximations are useful in designing measurements and in studying relationships between traffic measurement errors and the performance of the provisioning and administration processes. For example, the relationship of actual traffic measurement accuracies (which can be further corrupted by wiring, data base, and recording errors) to the quality of the trunk provisioning process was studied in Ref. 1. The variance approximation developed here was useful in quantifying the background accuracy of the process.

This paper is organized as follows. The basic approximation is presented and discussed in Section II. The development of the approximation is given in Section III; supporting analysis of switch count error is developed in the appendix. Concluding remarks are given in Section IV.

II. THE BASIC APPROXIMATION

Figure 1 illustrates UPCO measurements for a measurement interval of length T, with usage scan interval s. The UPCO estimate for the offered load during this measurement interval is given by



PEG COUNT $P = \Sigma$ (OFFERED CALLS IN MEASUREMENT INTERVAL)

OVERFLOW $O = \Sigma$ (OVERFLOW CALLS IN MEASUREMENT INTERVAL)

USAGE $$U = \frac{1}{\left(\frac{\text{NO. SCANS IN}}{\text{MEASUREMENT INTERVAL}}\right)\text{SCANS}} \Sigma \text{ (NO. TRUNKS BUSY)}$$

Fig. 1—UPCO measurements.

$$\hat{a} = \frac{\text{average measured usage}}{\text{1-measured blocking}}, \tag{1}$$

where the measured blocking is the ratio of overflow to offered attempts. It is well known that (under reasonable conditions subsequently discussed) this is an unbiased estimate for the true offered load a during this interval.

Early work on analyzing offered load estimators was carried out, among others, by R. I. Wilkinson,[2] who addressed the reliability of holding time estimates. In a 1952 paper,[3] W. S. Hayward, Jr., drawing on some of Wilkinson's analysis, addressed the variance of offered load estimates based on sampled usage. Hayward's model assumed Poisson arrivals, exponential holding times, and no blocking, yielding the result

$$\text{var}(\hat{a}) = \frac{\bar{h}a}{T}(2 + q), \tag{2}$$

where a is the offered load in erlangs, $\bar{h}$ is the average holding time, and T is the length of the measurement interval. The parameter q is given by

$$q = v\frac{1 + e^{-v}}{1 - e^{-v}} - 2, \tag{3}$$

where $v = s/\bar{h}$, and s is the usage scan interval; q determines the variance contribution due to switch count (sampling) error, e.g., $q = 0$ for $s = 0$, the continuous scan case.

In more recent work, Hill and Neal[4] addressed the question of the variance of $\hat{a}$ for peaked traffic,* but did not consider congestion or switch count error. Through the application of an asymptotic result for the variance of the renewals for a peaked traffic stream, they obtained the expression

$$\text{var}(\hat{a}) \cong \frac{2\bar{h}az}{T}, \tag{4}$$

where z is the peakedness factor for the stream.

In this paper, we combine elements of both of these previous analyses

* Peaked traffic refers to overflow traffic, or to streams containing some overflow traffic. The peakedness factor $z(\mu)$ (or z if μ is understood) is the equilibrium variance-to-mean ratio of busy servers when this traffic is offered to an infinitely large group of exponential servers with service rate μ. The peakedness factor is one for Poisson traffic and is larger than one for overflow traffic.

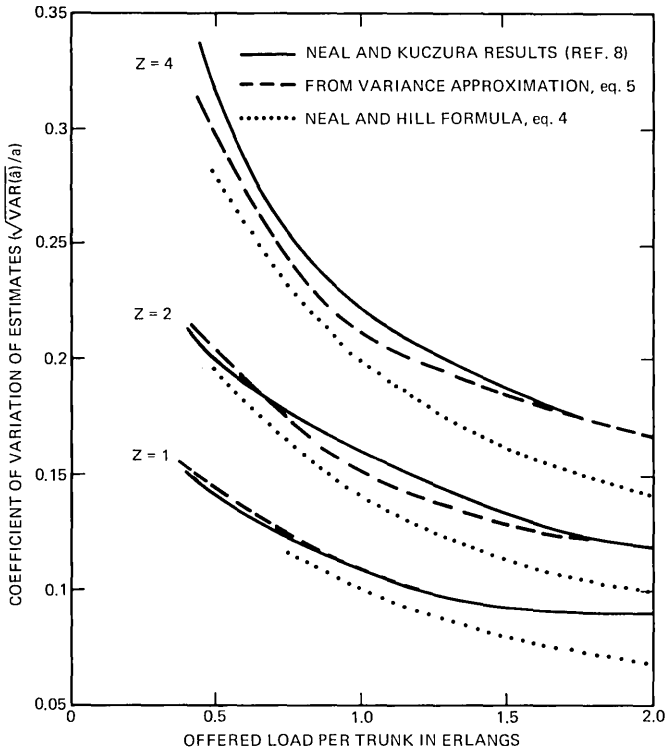


Fig. 2—Comparison of variance approximations for $N = 10$ servers ($\bar{h} = 180$ s, $T = 3600$ s, $s = 100$ s).

and explicitly consider the effect of blocking on the group, to obtain the generalization

$$\text{var}(\hat{a}) \cong \frac{\bar{h}a}{T}\left(2z + \frac{B+q}{1-B}\right), \tag{5}$$

where B is the equilibrium call congestion,* i.e., the fraction of attempts blocked. Thus, congestion basically adds a term to the previous various approximations.

Figures 2 and 3 show comparisons of the variance approximation (5) with the reference approximations obtained via the error theory devel-

* The blocking B is defined in theory as the probability that an arbitrary attempt is blocked. In practice, when the load parameters a,z are given, the blocking or call congestion B is assumed to be defined by the equivalent random method (Ref. 5), so that $B = f(N,a,z)$ where N is the number of trunks in the group. Otherwise, as shown by Holtzman (Ref. 6), the blocking B is not uniquely defined by N,a,z, but may take on a range of values, depending on higher order characteristics of the traffic stream. The actual value of $f(N,a,z)$ may be obtained from traffic tables normally used in administering trunking networks. It may also be estimated by Hayward's approximation, $f(N,a,z) \cong B(N/z,a/z)$ (Ref. 7), thus allowing Erlang $B(.,.)$ tables or formulas to be used.

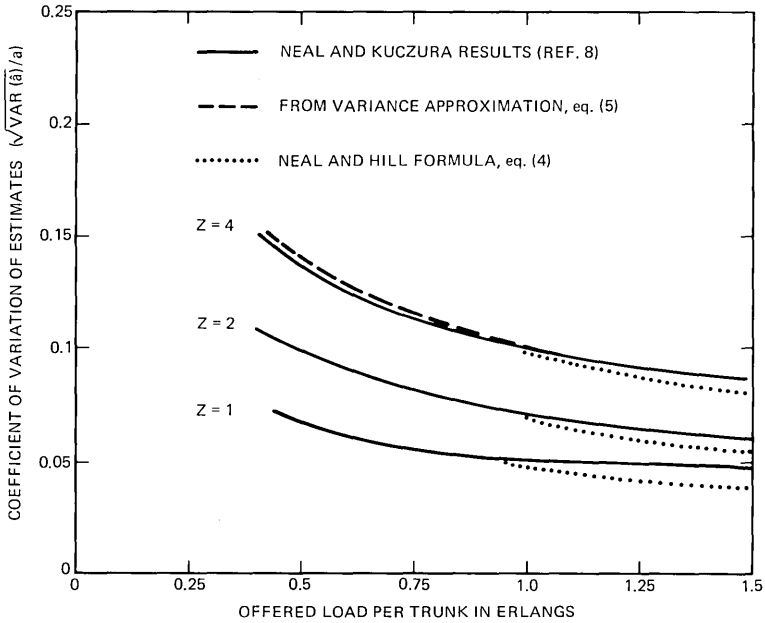


Fig. 3—Comparison of variance approximations for $N = 40$ servers ($\bar{h} = 180$ s, $T = 3600$ s, $s = 100$ s).

oped by Neal and Kuczura.[8*] These results assume that $\bar{h} = 180$ s, $T = 3600$ s (i.e., $\bar{h}/T = 0.05$), and $s = 100$ s. For a wide range of congestion and peakedness conditions, the agreement between eq. (5) and the reference results is very good. Neal and Kuczura also determined by numerical comparisons that switch count error was a small contributor to $\text{var}(\hat{a})$. Since q is small for typical scan-interval-to-holding-time ratios (e.g., $q \cong 0.05$ for $s = 100$ s and $\bar{h} = 180$ s, which are typical scan intervals and holding times for Bell System trunks), this conclusion is also evident from eq. (5).

Figures 2 and 3 also show the behavior of the Neal and Hill result, eq. (4). As the load per trunk increases, it is clear that the contribution of the congestion term in eq. (5) is increasingly important. These higher levels of congestion occur quite commonly on high usage groups, where a substantial fraction of the busy hour loads may be overflowed to an alternate route. As the load is increased to very large values, the coefficient of variation using eq. (4) goes to 0, whereas Figs. 2 and 3 suggest that the coefficient of variation has a positive limit as $a \to \infty$. It can be shown that (for any z) as the attempt rate $\lambda \to \infty$,

$$\lim_{\lambda \to \infty} \text{var}(\hat{a})/a^2 = \bar{h}/TN, \tag{6}$$

* This error theory is applicable to general functions of the e UPCO measurements. The approximation developed for the UPCO offered load estimate is computationally much more complex, as well as less transparent, than eq. (5). The Neal and Kuczura approximation agreed well with simulation results, and hence is a suitable reference for comparing eq. (5).

where N is the number of servers in the group.* Equation (6) has a simple interpretation. The UPCO offered load estimate may be viewed as the product of essentially independent estimators for the attempt rate λ and for the mean holding time $\bar{h}$. As $\lambda \to \infty$, the coefficient of variation for the first estimator goes to 0. Equation (6) represents the squared coefficient of variation for the second estimator, i.e., the positive limit results from having only a finite number of carried attempts from which to estimate mean holding time. For Figs. 2 and 3, the asymptotic limits for the coefficient of variation are 0.071 and 0.035, respectively.

If a is assumed to have a mean a_o and variance $\sigma_a{}^2$, one is often interested in estimating a_o. The results of this section can be applied to obtain $\text{var}(\hat{a}_o)$ for a single measurement period by interpreting them as conditional results, i.e., $\text{var}(\hat{a}\,|\,a)$, in the expression

$$\text{var}(\hat{a}_o) = \sigma_a{}^2 + E_a \text{ var}(\hat{a}\,|\,a). \tag{7}$$

In many cases, the $\sigma_a{}^2$ term can be a significant contributor. For example, in trunk engineering $\sigma_a{}^2$ may represent a day-to-day variance under an i.i.d. model for busy-hour loads (in this case, a_o is usually estimated from 5 to 20 busy-hour loads) and can be quite large in relation to the other sources of variability.

III. DEVELOPMENT OF THE APPROXIMATION

Consider a full access group of N servers operated in a blocked-calls-cleared mode. The offered traffic process is assumed to be a (nonlattice) renewal process with rate parameter λ, and server holding times are assumed to be exponential with hang-up rate μ. We define the mean and peakedness of the offered load by $a = \lambda/\mu$, $z = \text{var}(n(t))/E(n(t))$, where $n(t)$ is the equilibrium occupancy when the renewal process is offered to an infinitely large group of exponential servers with rate μ. The parameters (a,z) are conventionally used in traffic engineering, and hence it is useful to relate the variance approximation to these parameters.

For a measurement period of length T, let u,p,o denote average measured usage, offered attempts, and overflow attempts, as illustrated by Fig. 1. The average measured usage is defined by $u = 1/m \sum_{j=1}^{m=[T/s]} n(js)$ if $s > 0$, and by $u = 1/T \int_0^T n(t)dt$ if $s = 0$, where $n(t)$ is the number of busy servers at time t. It is assumed that equilibrium conditions apply at the beginning of the measurement interval, both for the occupancy on the servers and for the renewal processes corresponding to arrivals and overflows.

* This result is not the same as the limit obtained from eq. (5) as $a \to \infty$, which gives $(1 + q)\bar{h}/TN$. The discrepancy arises because the model for switch count error used in the development of eq. (5) breaks down as $\lambda \to \infty$. For this unrealistic limiting case, the servers are occupied 100 percent of the time, and *no* error is introduced by scanning. The correct result is thus obtained by noting that the carried attempt process approaches a Poisson process with rate $N\mu$ as $\lambda \to \infty$.

The UPCO estimate for the offered load a over the measurement period is

$$\hat{a} = \frac{u}{1 - o/p} = p\frac{u}{c} = \frac{p}{T}\left(\frac{Tu}{c}\right), \tag{8}$$

where $c \triangleq p - o$. Thus, $\hat{a}$ may be viewed as the product of separate estimators for the arrival rate (p/T) and for the average holding time (Tu/c). The approximation for $\text{var}(\hat{a})$ is obtained by introducing an approximate treatment of the scanning error, and then by examining (8) for large T. However, while the *structure* of the approximation is motivated by asymptotic analysis, the *validity* of the approximation is based on its accuracy for realistic values of T.

3.1 Treatment of scanning error

The scanning error for usage affects only the value Tu in (8), which may be expressed as

$$Tu = \sum_{j=1}^{c} \hat{h}_j + r_0 - r_T, \tag{9}$$

where $\hat{h}_j$ is the sampled holding time estimate for the jth call to be accepted by the group, $\hat{h}_j \in \{0,s,2s,\cdots\}$, and r_0, r_T are end effects. In particular, if the jth call to be accepted by the group was hit by k_j scans, then $\hat{h}_j = k_j s$ may be viewed as the sampled holding time estimate for this call. The variable r_0 is the total measurement period usage attributable to calls already in progress at the beginning of the interval, while r_T is the total usage due to accepted calls that would be measured in the subsequent measurement period of length T.

Throughout this analysis, we make the following simplifying assumptions:

(*i*) $\hat{h}_j, j = 1,2,\cdots,c$ are independent random variables.

(*ii*) $\hat{h}_j = h_j + e_j$ where e_j is the scanning error that results when a call with exponential holding time h_j begins at a time which is uniformly distributed between two successive sampling instants.

These simplifying assumptions hold *exactly* for the case $B = 0$, $s = 0$ (no congestion and continuous scan) and any z, since all calls are carried and the holding times are i.i.d. exponential random variables. They also hold *exactly* for the case $B = 0$, $s > 0$, and $z = 1$, since for a Poisson process the arrivals in disjoint intervals are independent. Furthermore, given a *fixed* number of arrivals in an interval (in particular, an interval of length s), the arrival times are independent and uniformly distributed within the interval. Thus, the simplifying assumptions—while not always true—can be rigorously justified for some important cases. In general, they can be expected to be reasonable assumptions if the usage on each server in the group does not approach unity, i.e., if congestion is not too severe.

As a result of the simplifying assumptions, the scanning error need only be examined for an isolated call. The analysis for this situation is treated in the appendix, where it is shown that with $e = \hat{h} - h = ks - h$, i.e., the sampled holding time minus the true holding time,

$$E(e) = 0 \tag{10}$$

$$\text{cov}(h,e) = 0 \tag{11}$$

$$\text{var}(e) = \bar{h}^2 \left[\upsilon \frac{1 + e^{-\upsilon}}{1 - e^{-\upsilon}} - 2 \right] \triangleq \bar{h}^2 q, \tag{12}$$

where $\upsilon = s/\bar{h}$, $\bar{h} = \mu^{-1}$. For $s = 0$ (continuous scan), $\text{var}(e) = 0$ as expected, and hence these results cover both the continuous or the discrete scan case.

3.2 Asymptotic analysis of variance

Since p corresponds to the arrivals for a renewal process, $x \triangleq p/T$ is asymptotically normal with mean λ and variance of the form $O(1/T)$ (Ref. 9, p. 40). It is established in Ref. 10 that the variance can be approximately expressed in terms of the peakedness z

$$\text{var}(x) \cong (2z - 1)\lambda/T. \tag{13}$$

As noted in Ref. 4, this approximation has been found to be quite good for $a > z - 1$, and $T \geqq 10\bar{h}$. Although the carried calls c do not necessarily correspond to a renewal process (unless $c \equiv p$), c/T is also asymptotically normal with mean $\lambda(1 - B)$, (where $B \triangleq \lim_{T \to \infty}(o/p)$) and variance $O(1/T)$. This follows since if $B > 0$ the overflow process o is a renewal process, and the carried calls between overflows are independent for successive interoverflow periods. The only other asymptotic result needed is the following one, the proof of which is essentially the same as that for the function of sampling moments theorem given on p. 366 of Cramér:[11]

If $g(.,.)$ is a twice continuously differentiable function in some neighborhood of the point $\lambda,\lambda(1 - B)$, then $g(p/T,c/T)$ is asymptotically normal with mean $g(\lambda,\lambda(1 - B))$ and variance $O(1/T)$. It follows that

$$E(g(p/T,c/T)) = g(\lambda,\lambda(1 - B)) + O(1/\sqrt{T}). \tag{14}$$

Now for large T, the end effects r_0,r_T in (9) can be ignored at the outset. In particular, we have $E(Tu) = O(T)$, $\text{var}(Tu) = O(T)$, whereas $E(r_0 - r_T) = o(1)$, $\text{var}(r_0 - r_T) = O(1)$. (In general, ignoring these end effects is valid when $T/\bar{h}$ is reasonably large, e.g., $T/\bar{h} \geqq 10$.) Thus, defining* $y = \Sigma_{j=1}^{c} \hat{h}_j/c$, where the $\hat{h}_j$ satisfy the simplifying assumptions made for handling the scanning error, it follows from (10) to (12) that

* While y can be defined to be 0 for $c = 0$, in order to simplify subsequent notation, we shall assume that $P(c = 0) = 0$. This is reasonable even for the typical values of T that are of interest in practical applications.

$$E(y) = E(\hat{h}) = \bar{h} \tag{15}$$

$$\text{var}(y) = \text{var}(\hat{h})E\left(\frac{1}{c}\right) = \frac{\bar{h}^2(1+q)}{\lambda T(1-B)} + o(1/T), \tag{16}$$

where we have used (14) to evaluate $E(1/c)$.

Turning our attention next to $\hat{a}$, we have

$$\hat{a} = xy \tag{17}$$

$$\text{var}(\hat{a}) = E(x^2y^2) - E^2(xy). \tag{18}$$

In order to simplify this expression, we first note that

$$E(y|c) = \bar{h}$$

and hence

$$E(xy) = E_{p,c}E(xy|p,c) = E_{p,c}(x\bar{h}) = \lambda\bar{h} = E(x)E(y); \tag{19}$$

i.e., x,y are uncorrelated, confirming that $\hat{a}$ is an unbiased estimate of a. By the same conditioning, we also obtain

$$E(x^2y^2) = \bar{h}^2(E(x^2) + (1+q)E(x^2/c)) \tag{20}$$

and since

$$E(x^2)E(y^2) = \bar{h}^2(E(x^2) + (1+q)E(1/c)E(x^2)), \tag{21}$$

$$E(x^2y^2) = E(x^2)E(y^2) + (1+q)\bar{h}^2w, \tag{22}$$

where $w = \text{cov}(x^2,1/c)$. Substituting (19) and (22) into (18) and identifying terms, we have

$$\text{var}(\hat{a}) = E^2(x)\,\text{var}(y) + E^2(y)\,\text{var}(x) + \text{var}(x)\,\text{var}(y) + (1+q)\bar{h}^2w. \tag{23}$$

By direct substitution of the means and variances for x,y

$$\text{var}(\hat{a}) = \frac{a\bar{h}(1+q)}{T(1-B)} + (2z-1)\frac{a\bar{h}}{T} + o(1/T) + (1+q)\bar{h}^2w. \tag{24}$$

It remains to show that $w = o(1/T)$. But $Tw = \text{cov}(x^2,1/(c/T))$ and hence by (14) it follows that $Tw = o(1)$, i.e., $w = o(1/T)$. This completes the analysis; the variance approximation given in eq. (5) corresponds to terms of $O(1/T)$ in (24).

IV. CONCLUSIONS

In this paper, we have developed a simple approximation for the variance of the UPCO offered load estimate commonly used in offered load estimation. This approximation shows clearly the role of source load variation, switch count error, peakedness, congestion, and length of the measurement period. Relative to previous work, the main contribution is the explicit inclusion of congestion. Thus the results are of particular

interest for high congestion situations such as occur in measuring loads on high usage groups.

While the basic approximation is developed here for a single measurement interval, it can be easily applied in analyzing load estimates based on the average load over a number of single measurement intervals.

V. ACKNOWLEDGMENTS

Discussions with S. R. Neal and D. W. Hill and D. L. Jagerman are gratefully acknowledged.

APPENDIX

Analysis of Switch Count Error

In this appendix we analyze, using methods similar to Hayward,[3] the following switch count error model: (*i*) a call with holding time h begins at a time uniformly distributed between two successive sampling instants, (*ii*) the sampling interval is of length s, (*iii*) the holding time is exponentially distributed with rate parameter μ.

For an arbitrary call, the error e between the true holding time h for the call, and the "sampled holding time," is given by $e = ks - h$, where k represents the scan count for the call, $k \epsilon \{0,1,2,\cdots\}$. The scan count for the call is simply the total number of scans that occur during the time the call is in progress.

Since $e \epsilon [-s,s]$, it is convenient to define a normalized error $e' = k - h'$, where $h' = h/s$ is exponentially distributed with rate parameter $\mu' = \mu s = s/\overline{h}$. The density of h' is therefore given by

$$f(t) = \begin{cases} 0 & t < 0 \\ \mu' e^{-\mu' t} & t \geq 0 \end{cases}. \tag{25}$$

Define $x' = x/s$, where x is uniformly distributed in $[0,s]$ and represents the time from a sampling instant to the beginning of a call. Given $x' \epsilon [0,1]$, it is straightforward to show that the conditional probability density of e' at $e' = y$ is

$$g(y|x') = \begin{cases} 0, & y \notin [-(1-x'),x'] \\ \sum_{k=0}^{\infty} f(k-y), & y \epsilon [-(1-x'),x']. \end{cases} \tag{26}$$

The only case for which a negative argument can occur in any term in the preceding sum is for $k = 0, y > 0$. Thus,

$$\sum_{k=0}^{\infty} f(k-y) = \frac{e^{-\mu'}}{1-e^{-\mu'}} \mu' e^{\mu' y} \text{ for } y > 0$$

$$\sum_{k=0}^{\infty} f(k-y) = \frac{1}{1-e^{-\mu'}} \mu' e^{\mu' y} \text{ for } y < 0.$$

Defining $r = e^{-\mu'}$, (26) becomes

$$g(y|x') = \begin{cases} 0 & x' < y \le 1 \\ \dfrac{r}{1-r}\mu' e^{\mu' y} & 0 \le y \le x' \\ \dfrac{1}{1-r}\mu' e^{\mu' y} & -(1-x') \le y < 0 \\ 0 & -1 \le y < -(1-x'). \end{cases} \tag{27}$$

To simplify obtaining of moments for e', we define $G(\alpha) = E(e^{\alpha e'}) = E_{x'}E(e^{\alpha e'}|x')$. Using (27),

$$G(\alpha) = \frac{1}{1-r} E_{x'}\left[r \int_0^{x'} \mu' e^{(\mu'+\alpha)y} dy + \int_{-(1-x')}^{0} \mu' e^{(\mu'+\alpha)y} dy \right]. \tag{28}$$

After integration, one obtains

$$G(\alpha) = \left(\frac{\mu'}{\mu'+\alpha}\right) - \left(\frac{1+r}{1-r}\right)\frac{\mu'}{(\mu'+\alpha)^2} + \frac{(e^{\alpha} + re^{-\alpha})}{1-r}\frac{\mu'}{(\mu'+\alpha)^2}. \tag{29}$$

We have $G(0) = 1$, $G'(0) = 0$, and

$$G''(0) = \frac{1+r}{1-r}\frac{1}{\mu'} - 2\frac{1}{(\mu')^2}; \tag{30}$$

hence,

$$E(e) = 0 \tag{31}$$

$$\operatorname{var}(e) = \bar{h}^2\left(\frac{1+e^{-s/\bar{h}}}{1-e^{-s/\bar{h}}}\cdot\frac{s}{\bar{h}} - 2\right), \tag{32}$$

which establishes (10) and (12) of the main section.

To establish the covariance between h, e, we note that because of (31), $\operatorname{cov}(h,e) = E(h\,e) = s^2E(h'e')$. But

$$\begin{aligned} E(h'e') &= E_{x'}\left[\int_{-(1-x')}^{x'} \sum_{k=0}^{\infty} y(k-y)f(k-y)dy \right] \\ &= E_{x'}\left[\int_{-(1-x')}^{x'} \sum_{k=0}^{\infty} (-y^2)f(k-y)dy \right] \\ &\quad + E_{x'}\left[\int_{-(1-x')}^{x'} \sum_{k=0}^{\infty} kyf(k-y)dy \right]. \end{aligned} \tag{33}$$

The first term is $-\operatorname{var}(e')$. To evaluate the second term, we note that

$$\sum_{k=0}^{\infty} kyf(k-y) = \sum_{k=0}^{\infty} ky\mu' e^{-\mu'(k-y)} = ye^{\mu' y}\mu' \sum_{k=0}^{\infty} kr^k$$

$$= ye^{\mu' y}\mu' r \sum_{k=1}^{\infty} kr^{k-1} = ye^{\mu' y}\mu' r \frac{d}{dr}\left(\frac{1}{1-r}\right), r = e^{-\mu'}.$$

Therefore

$$E_{x'}\left[\int_{-(1-x')}^{x'} \sum_{k=0}^{\infty} kyf(k-y)dy\right]$$

$$= \frac{r}{(1-r)^2} E_{x'}\left[\int_{-(1-x')}^{x'} y\mu' e^{\mu' y} dy\right]. \quad (34)$$

Thus, we are led to define the function

$$H(\alpha) = E_{x'}\left[\int_{-(1-x')}^{x'} \mu' e^{(\mu+\alpha)y} dy\right].$$

Carrying out the integration yields

$$H(\alpha) = -\frac{2\mu'}{(\mu'+\alpha)^2} + \frac{\mu'}{(\mu'+\alpha)^2} \frac{e^{\alpha} + r^2 e^{-\alpha}}{r}. \quad (35)$$

The expectation in (34) is now evaluated as

$$H'(0) = \frac{1}{\mu'} \frac{(1-r)(1+r)}{r} + \frac{1}{(\mu')^2}\left(4 - 2\frac{1+r^2}{r}\right),$$

giving

$$E_{x'}\left[\int_{-(1-x')}^{x'} \sum_{k=0}^{\infty} kyf(k-y)dy\right] = \frac{1}{\mu'}\left(\frac{1+r}{1-r}\right) - \frac{2}{(\mu')^2} = \text{var}(e').$$

Therefore, $E(h'e') = -\text{var}(e') + \text{var}(e') = 0$, i.e., h' and e' are uncorrelated random variables and

$$\text{var}(\hat{h}) = \text{var}(ks) = \text{var}(h) + \text{var}(e). \quad (36)$$

Remark: Hayward[3] treats switch count error and source load variation separately, assumes independence, and adds the separate variances to obtain an approximate result. He noted that the errors were probably correlated, though weakly, and that (at that time) no method to take this into account was evident (Ref. 3, p. 363). Since $\text{cov}(h,e) = 0$, it follows from this analysis that (for the same model studied by Hayward) the errors are in fact uncorrelated. It was also pointed out by the referee that an alternate proof that $\text{cov}(h,e) = 0$ can be obtained by noting that the scan count k is geometrically distributed for $k \geq 1$. Thus, by directly evaluating $\text{var}(ks)$, one finds that $\text{var}(ks) = \text{var}(h) + \text{var}(e)$, which implies $\text{cov}(h,e) = 0$.

REFERENCES

1. R. L. Franks, H. Heffes, J. M. Holtzman, and S. Horing, "A Model Relating Measurement and Forecast Errors to the Provision of Direct Final Trunk Groups," Proceedings of the 8th ITC (November 1976), pp. 133-1 to 133-7.
2. R. I. Wilkinson, "The Reliability of Holding Time Measurements," B.S.T.J., *20,* No. 4 (October 1941), pp. 365–404.
3. W. S. Hayward, Jr., "The Reliability of Telephone Traffic Load Measurements by Switch Counts," B.S.T.J., *31,* No. 2 (March 1952), pp. 357–377.
4. D. W. Hill and S. R. Neal, "Traffic Capacity of a Probability Engineered Group," B.S.T.J., *55,* No. 7 (September 1976), pp. 831–842.
5. R. I. Wilkinson, "Theories for Toll Traffic Engineering in the U.S.A.," *BSTJ, 35,* No. 2 (March 1956), pp. 421–514.
6. J. M. Holtzman, "The Accuracy of the Equivalent Random Method with Renewal Inputs," B.S.T.J., *52,* No. 9 (November 1973), pp. 1673–1679.
7. W. S. Hayward, Jr., unpublished work.
8. S. R. Neal and A. Kuczura, "A Theory of Traffic Measurement Errors for Loss Systems with Renewal Input," B.S.T.J., *52,* No. 6 (July–August 1973), pp. 967–990.
9. D. R. Cox, *Renewal Theory,* London: Methuen, and New York: Wiley, 1962.
10. D. L. Jagerman, unpublished work.
11. Harald Cramér, *Mathematical Methods of Statistics,* Princeton: Princeton University Press, 1946.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 7, September 1978
Printed in U.S.A.

Adaptive Equalization of Channel Nonlinearities in QAM Data Transmission Systems

By D. D. FALCONER

(Manuscript received August 23, 1977)

Within the population of voiceband telephone channels, few channel characteristics are as pervasive in their impairment of high-speed data communication as nonlinear distortion, which cannot be removed or equalized in the receiver as easily as can linear distortion. The purpose of this paper is to report on an investigation of a QAM receiver incorporating adaptive equalization of nonlinearities as well as adaptive decision feedback equalization and data-aided carrier recovery for mitigation of linear distortion and phase jitter, respectively. Nonlinearities are equalized by adding to the received in-phase and quadrature signals a weighted sum of nonlinear functionals of the received signal and of modulated previous receiver decisions. The choice of nonlinear terms in the sum is based on a channel model incorporating quadratic and cubic nonlinearities as well as linear dispersive elements. The adjustment of the weighting, or tap, coefficients for the various terms is based on a gradient algorithm, as is the adjustment of the linear tap coefficients and the carrier phase reference. The feasibility of nonlinearity equalization on real voiceband channels was confirmed in a test in which recorded 9600-bps QAM signals, received from a worse-than-average set of 17 voiceband telephone channels, were processed by a computer-simulated version of the proposed receiver (termed the NL receiver). The observed error rates for all channels were lower, in some cases by several orders of magnitude, than those achieved by computer-simulated versions of the linear receiver and of a decision feedback equalization receiver (termed the DFE receiver).

I. INTRODUCTION

The prevalence of nonlinearities and their distorting effect on high-speed data transmission over voiceband telephone channels has long been recognized.[1] The effect of nonlinear distortion on linearly modulated data signals is to introduce nonlinear intersymbol interference and

reduce the margin against noise. For data rates above 4800 bps, nonlinear distortion is the dominant impairment on many voiceband telephone channels. Experimental studies have measured nonlinear distortion and related the observed error rates for specific modulation formats to this and other measured impairments.[2,3] Estimation of performance for data transmission in the presence of nonlinearities can be carried out[4] but gives little insight into the problem of receiver optimization, except for certain simple nonlinear channel models.[5]

Recognizing that nonlinearities in transmission channels generally coexist with linear elements such as filters, one is led to consider a general nonlinear receiver structure, based on a Volterra or Wiener kernel characterization[6] of a general nonlinear system such as that proposed in Refs. 5, 7, and 8, the latter in connection with adaptive echo cancellation. In the present work, we extend this approach by generalizing the structure of a passband decision feedback equalizer, previously studied in connection with linear channel distortion,[9] to process nonlinear as well as linear functionals of the incoming signal and prior decisions.*

The new receiver structure is based on a model of a passband channel with quadratic and cubic nonlinearities, as well as linear filters. We report on the simulation of the new receiver and on comparisons of its performance with two other previously simulated 9600-bps QAM receivers on a worse-than-average set of voiceband telephone channels. The new receiver is referred to as the NL receiver. The other two receivers, designated LE (linear equalization) and DFE (decision feedback equalization), are not designed to compensate for channel nonlinearities. Their performance is compared over the same set of voiceband telephone channels in Ref. 9. The simulated LE receiver is described in Ref. 10.

II. SUMMARY OF THE MAJOR RESULTS

The relative performances of the three simulated receivers on the same set of recorded, received, 9600-bps data signals are briefly summarized as follows: On every channel, the NL receiver yielded a lower error probability than the other two receivers. For 13 out of the 17 channels, the improvement in error rate was equal to or better than about an order of magnitude. Another gauge of the degree of improvement offered by the NL receiver is the fact that it increased the number of channels yielding a better-than-10^{-4} error rate from 8 to 15. On one channel, whose major impairment was second harmonic distortion, the NL re-

* Figure 3a summarizes the structure of the nonlinearity-equalizing receiver.

ceiver's error rate bested that of the DFE and LE receiver by over four orders of magnitude. Figure 5 is a bar graph summarizing the error rate comparisons.

The apparent attractiveness of the NL receiver structure is, however, tempered by its greater complexity. A large number of nonlinear tap coefficients is necessary to account and compensate for the dispersive nonlinear effects typically encountered on voiceband channels. In the simulations summarized above, the LE and DFE receivers each had 32 complex tap coefficients, but the NL receiver was, roughly speaking, comparable in complexity to an LE receiver with 134 complex tap coefficients. Reducing the number of coefficients in the NL receiver lowered its performance margin over the other receivers. Furthermore, the best allocation of a fixed number of tap coefficients varied from one channel to another. These points are explored more fully in later sections.

In spite of the greater complexity of the NL receiver structure, the performance comparison of the three receivers does indicate the importance of alleviating nonlinear distortion for high-speed data transmission.

III. THE CHANNEL MODEL

Obviously, the effect of channel nonlinearities on a passband QAM data signal must be understood before a compensating receiver structure can be suggested. A general representation of a bandlimited QAM signal is as the real part of a complex waveform:

$$x(t) = \mathrm{Re}\left[e^{j2\pi f_c t} \sum_n A(n)F(t - nT) \right], \tag{1}$$

where $j = \sqrt{-1}$, f_c is the carrier frequency, $A(n)$ is a quantized complex number representing the information symbol in the nth symbol interval (for example, in the case of four-level QAM, the real and imaginary parts of $A(n)$ assume one of the four possible values $\pm 1, \pm 3$), T is the reciprocal of the baud, and $F(t)$ is a complex pulse waveform.

In the case of QAM signals, extraction of the information symbols represented by the complex number $A(n)$ requires two receiver outputs, which are derived by appropriate operations on both the received passband signal and on its quadrature version, or Hilbert transform. A phase-splitting filter is used to obtain both in-phase and quadrature versions of a voiceband data signal.

The complex waveform

$$X(t) = e^{j2\pi f_c t} \sum_n A(n)F(t - nT) \tag{2}$$

is assumed analytic;[11] that is, its spectrum is twice the Fourier transform of $x(t)$ for positive frequencies and is zero elsewhere. Furthermore, we assume the spectrum is limited on the high side to frequency $2f_c$. Note that the Fourier transform $\mathcal{F}(f)$ of the complex pulse $F(t)$ is not necessarily symmetric about $f = 0$, but it is assumed to be strictly band-limited to $-f_c < f < f_c$. The Nyquist frequency is $1/2T$ Hz. Figure 1 shows a sketch of $\mathcal{F}(f)$ and of $\mathcal{F}(f - f_c)$, which is the Fourier transform of $e^{j2\pi f_c t}F(t)$.

The notion of analytic signals is a notational convenience. The Hilbert transform, or quadrature version of a signal $u(t)$, is a linear functional of $u(t)$:

$$\check{u}(t) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{u(\tau)}{t - \tau}\, d\tau.$$

It can be shown that there is a unique analytic signal whose real part is $u(t)$, and that $\check{u}(t)$ is then just the imaginary part of the analytic signal. Conversely, any analytic signal comprises some real signal plus j times its Hilbert transform. Since QAM systems operate on both in-phase and quadrature versions of signals, they are most conveniently represented by means of analytic signals.

The nonlinear receiver structure will be based on the simple nonlinear channel model shown in Fig. 2, using the notation of analytic signals. Filters 1, 2, and 3 are passband with the same bandwidth as the transmitted data signal. The filters may include the receiver's input filter as well as the linear response of the channel. The quadratic and cubic memoryless nonlinearities with attenuated outputs account for second

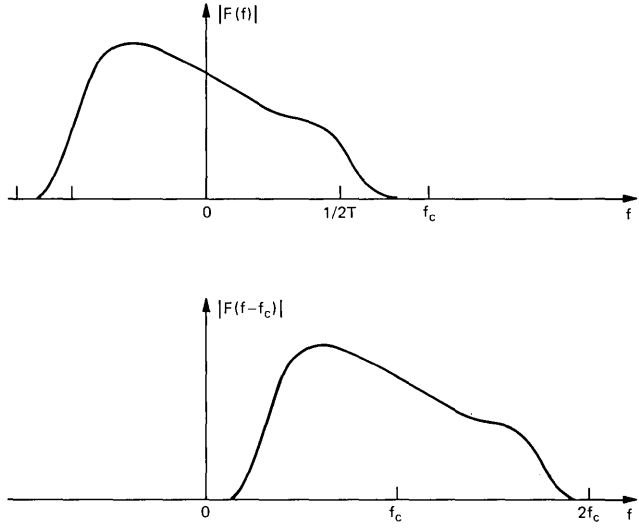


Fig. 1—Fourier transforms of $|F(f)|$ and $|F(f - c)|$.

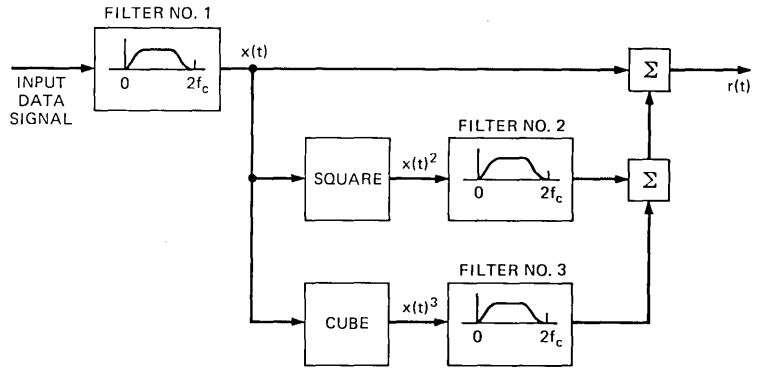


Fig. 2—Model of a nonlinear channel.

and third harmonic distortion, respectively. Additional impairments not shown in Fig. 2 are phase jitter, which implies multiplication of the complex received signal by $e^{j\phi(t)}$, and additive noise.

The result of passing the transmitted waveform through the linear portion of the channel (filter 1) is an analytic waveform in the form of eq. (2). A passband linear[10,12] equalizer (LE) can be used to minimize the mean squared error between its output, sampled at times nT, and a reference $A(n)e^{j(2\pi f_c nT+\hat{\theta}(n))}$, which is the complex information symbol modulated to passband with a receiver phase reference $\hat{\theta}(n)$. In a linear receiver, the passband equalizer output is demodulated [multiplied by $e^{-j(2\pi f_c nT+\hat{\theta}(n))}$] and then quantized to yield a decision $\hat{A}(n)$. A passband equalizer configuration which is theoretically more effective in combatting linear intersymbol interference is the passband DFE, described in Ref. 9.

To motivate a receiver structure which is appropriate for nonlinear distortion as well as linear distortion, we must consider the analytic signals emanating from the quadratic and cubic path elements of Fig. 2.

It is shown in the appendix that the analytic signal output from the model of Fig. 2 is of the form

$$R(t) = U_0(t) + e^{j2\pi f_c t}U_{11}(t) + e^{-j2\pi f_c t}U_{12}(t) + e^{j4\pi f_c t}U_2(t) + e^{j6\pi f_c t}U_3(t), \quad (3a)$$

where

$$U_0(t) = \sum_{n_1,n_2} A(n_1)A(n_2)^* G_0(t - n_1T, t - n_2T) \quad (3b)$$

$$U_{11}(t) = \sum_n A(n)F(t - nT) + \sum_{n_1,n_2,n_3} A(n_1)A(n_2)A(n_3)^* G_{11}(t - n_1T, t - n_2T, t - n_3T) \quad (3c)$$

$$U_{12}(t) = \sum_{n_1,n_2,n_3} A(n_1)^*A(n_2)^*A(n_3)G_{12}(t - n_1T, t - n_2T, t - n_3T) \tag{3d}$$

$$U_2(t) = \sum_{n_1,n_2} A(n_1)A(n_2)G_2(t - n_1T, t - n_2T) \tag{3e}$$

$$U_3(t) = \sum_{n_1,n_2,n_3} A(n_1)A(n_2)A(n_3)G_3(t - n_1T, t - n_2T, t - n_3T), \tag{3f}$$

where asterisks denote complex conjugates.

The various U terms are seen to be linear combinations of products of complex information symbols $A(n)$, $A(n_1)A(n_2)$, $A(n_1)A(n_2)A(n_3)^*$, etc. Each modulates a harmonic of the carrier wave. The term $e^{j2\pi f_c t}$-$U_{11}(t)$ includes the linear response of the channel to the data signal and also a component resulting from cubic distortion. The terms $U_0(t)$ and $U_2(t)$ result from the quadratic nonlinearity and the terms $U_{12}(t)$ and $U_3(t)$ result from the cubic nonlinearity. Additional terms would, of course, result from the assumption of additional nonlinear elements in the model of Fig. 2. The generalization of expression (3) to an infinite power series would be a complex passband version of a Volterra expansion.

IV. THE NONLINEAR RECEIVER STRUCTURE

The receiver structure to be studied here includes the passband QAM decision feedback equalizer discussed in Ref. 9, plus nonlinear processing suggested by the set of eqs. (3). Let $Y(n)$ be the receiver's complex output at time $t = nT$. This output is quantized to form the decision $\hat{A}(n)$, which equals the original transmitted symbol $A(n)$ if no error occurred. Let the demodulator's phase reference at time nT be $\hat{\theta}(n)$. Let $\{W_k^{(1)}\}_{k=-N}^{N}$ and $\{B_k^{(1)}\}_{k=1}^{M}$ be the complex linear forward and feedback tap coefficients respectively, and let $\{R(n)\}$ be the complex receiver input, sampled at times nT. Then

$$Y(n) = e^{-j(2\pi f_c nT + \hat{\theta}(n))} \sum_{k=-N}^{N} W_k^{(1)*} R(n-k) - \sum_{k=1}^{M} B_k^{(1)*} \hat{A}(n-k) + Y_{NL}(n)e^{-j(2\pi f_c nT + \hat{\theta}(n))}, \tag{4a}$$

where $Y_{NL}(n)$ consists of nonlinear functions of $\{R(k)\}$ and $\{\hat{A}(k)\}_{k<n}$.

The linear part of eq. (4a) implies a demodulated linear combination of $2N + 1$ receiver input samples minus a linear combination of M previous decisions.

The nonlinear term $Y_{NL}(n)$ is heuristically suggested by expression (3) in the following way: (*i*) Assume that at time nT the previous receiver decisions $\hat{A}(k) = A(k)(k < n)$ and that they are available to form the

nonlinear feedback terms. (*ii*) In any terms of expression (3) involving decisions $\hat{A}(k)$ not yet made at time $n (k \geq n)$, replace $\hat{A}(k)e^{j2\pi f_c kT+\hat{\theta}(k)}$ by $R(k)$ to form the forward nonlinear terms. The resulting expression is

$$
\begin{aligned}
Y_{NL}(n) = &\sum_{k_1,k_2} W^{(0)*}_{k_1k_2} R(n-k_1)R(n-k_2)* \\
&+ \sum_{k_1,k_2,k_3} W^{(11)*}_{k_1,k_2,k_3} R(n-k_1)R(n-k_2)R(n-k_3)* \\
&+ \sum_{k_1,k_2,k_3} W^{(12)*}_{k_1,k_2,k_3} R(n-k_1)*R(n-k_2)*R(n-k_3) \\
&+ \sum_{k_1,k_2} W^{(2)*}_{k_1,k_2} R(n-k_1)R(n-k_2) \\
&+ \sum_{k_1,k_2,k_3} W^{(3)*}_{k_1,k_2,k_3} R(n-k_1)R(n-k_2)R(n-k_3) \\
&- e^{j\hat{\theta}(n)} \sum_{\substack{k_1,k_2 \\ \geq 1}} B^{(0)*}_{k_1,k_2} \hat{A}(n-k_1)\hat{A}(n-k_2)* \\
&- e^{j(2\pi f_c nT+\hat{\theta}(n))} \sum_{\substack{k_1,k_2,k_3 \\ \geq 1}} B^{(11)*}_{k_1,k_2,k_3} \hat{A}(n-k_1)\hat{A}(n-k_2)\hat{A}(n-k_3)* \\
&- e^{-j(2\pi f_c nT-\hat{\theta}(n))} \sum_{\substack{k_1,k_2,k_3 \\ \geq 1}} B^{(12)*}_{k_1,k_2,k_3} \hat{A}(n-k_1)*\hat{A}(n-k_2)*\hat{A}(n-k) \\
&- e^{j(4\pi f_c nT+\hat{\theta}(n))} \sum_{\substack{k_1,k_2 \\ \geq 1}} B^{(2)*}_{k_1,k_2} \hat{A}(n-k_1)\hat{A}(n-k_2) \\
&- e^{j(6\pi f_c nT+\hat{\theta}(n))} \sum_{\substack{k_1,k_2,k_3 \\ \geq 1}} B^{(3)*}_{k_1,k_2,k_3} \\
&\qquad \times \hat{A}(n-k_1)\hat{A}(n-k_2)\hat{A}(n-k_3). \qquad (4b)
\end{aligned}
$$

The formidable-looking expression (4b) is a linear combination of products of receiver inputs and their complex conjugates, minus a linear combination of products of previous decisions and their complex conjugates, modulated by appropriate harmonics of the carrier.

Figures 3a and 3b are block diagrams of the NL receiver. The cross-hatched boxes in Figure 3a show the nonlinear processing that has been added to the basic decision feedback equalization structure described in an earlier paper.[9]

V. ADAPTATION OF RECEIVER PARAMETERS

As in the linear and decision feedback equalization receivers, the parameters $\{W\}$, $\{B\}$ and $\hat{\theta}$ are adjusted in an estimated gradient algorithm to minimize the average value of the squared error magnitude $|E(n)|^2$ defined by

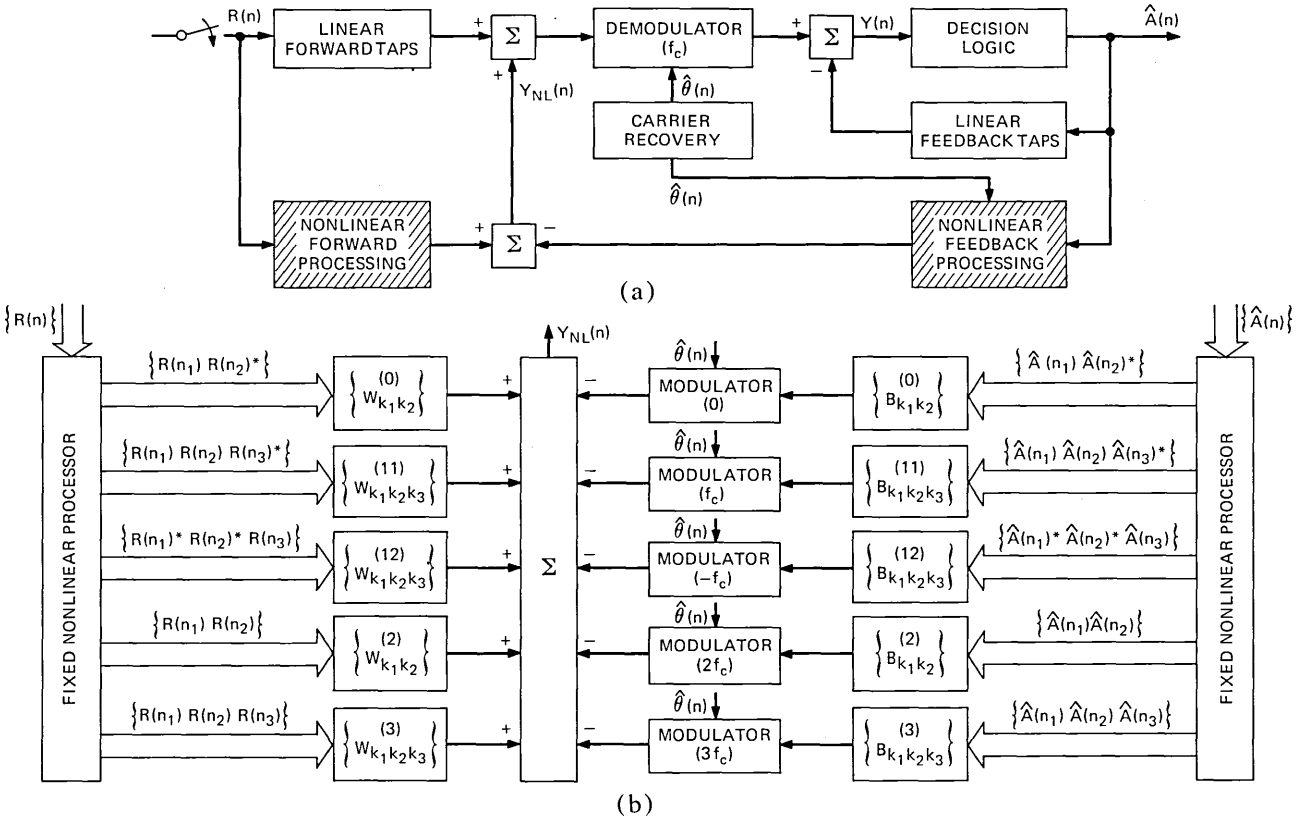


Fig. 3—(a) Basic structure of the NL receiver. (b) Details of the nonlinear signal processing.

$$E(n) \equiv Y(n) - A(n). \tag{5}$$

The error $E(n)$, as in the previous receivers, is a *linear* function of the parameters $\{W\}$ and $\{B\}$; consequently, the expression for $|E(n)|^2$ is convex in these parameters.

In writing the updating equations for the $\{W(n)\}$ and $\{B(n)\}$ coefficients and for $\hat{\theta}(n)$ in the nth symbol interval, it is convenient to use the symbol $\epsilon(n)$ to denote the *observed passband* error after the decision $A(n)$ has been made:

$$\epsilon(n) = [Y(n) - \hat{A}(n)]e^{j(2\pi f_c nT + \hat{\theta}(n))}; \tag{6a}$$

thus, if $\hat{A}(n) = A(n)$, $|E(n)|^2 = |\epsilon(n)|^2$, and the expression for the gradient of $|\epsilon(n)|^2$ with respect to each parameter determines an adjustment algorithm for that parameter. The adjustment equation for $\hat{\theta}(n)$ is as follows:

$$\hat{\theta}(n+1) = \hat{\theta}(n) - \frac{\alpha Im[\epsilon(n)^* Z(n)]}{|\hat{A}(n)|^2}, \tag{6b}$$

where
$$\begin{aligned} Z(n) = &\sum_{k=-N}^{N} W_k^{(1)*} R(n-k) \\ &+ \sum_{k_1,k_2} W_{k_1,k_2}^{(0)*} R(n-k_1)R(n-k_2)^* \\ &+ \sum_{k_1,k_2,k_3} W_{k_1,k_2,k_3}^{(11)*} R(n-k_1)R(n-k_2)R(n-k_3)^* \\ &+ \sum_{k_1,k_2,k_3} W_{k_1,k_2,k_3}^{(12)*} R(n-k_1)^* R(n-k_2)^* R(n-k_3) \\ &+ \sum_{k_1,k_2} W_{k_1,k_2}^{(2)*} R(n-k_1)R(n-k_2) \\ &+ \sum_{k_1,k_2,k_3} W_{k_1,k_2,k_3}^{(3)*} R(n-k_1)R(n-k_2)R(n-k_3) \end{aligned} \tag{6c}$$

is the sum of all the forward terms comprising $Y(n)$. The adjustment equations for the $\{W\}$ and $\{B\}$ coefficients are as follows:

$$W_{k_1,k_2}^{(0)}(n+1) = W_{k_1,k_2}^{(0)}(n) - \beta_0 \epsilon(n)^* R(n-k_1)R(n-k_2)^* \tag{6d}$$

$$W_k^{(1)}(n+1) = W_k^{(1)}(n) - \beta_1 \epsilon(n)^* R(n-k) \tag{6e}$$

$$W_{k_1,k_2,k_2}^{(11)}(n+1) = W_{k_1,k_2,k_3}^{(11)}(n) - \beta_{11}\epsilon(n)^* R(n-k_1) \cdot R(n-k_2)R(n-k_3)^* \tag{6f}$$

$$W_{k_1,k_2,k_3}^{(12)}(n+1) = W_{k_1,k_2,k_3}^{(12)}(n) - \beta_{12}\epsilon(n)^* R(n-k_1)^* \cdot R(n-k_2)^* R(n-k_3) \tag{6g}$$

$$W_{k_1,k_2}^{(2)}(n+1) = W_{k_1,k_2}^{(2)}(n) - \beta_2 \epsilon(n)^* R(n-k_1)R(n-k_2) \tag{6h}$$

$$W_{k_1,k_2,k_3}^{(3)}(n+1) = W_{k_1,k_2,k_3}^{(3)}(n) - \beta_3 \epsilon(n)^* R(n-k_1) \cdot R(n-k_2)R(n-k_3) \tag{6i}$$

$$B^{(0)}_{k1,k2}(n+1) = B^{(0)}_{k1,k2}(n) + \gamma_0 \epsilon(n)^* \hat{A}(n-k_1)\hat{A}(n-k_2)^* e^{j\hat{\theta}(n)} \quad (6j)$$

$$B^{(1)}_{k}(n+1) = B^{(1)}_{k}(n) + \gamma_1 \epsilon(n)^* A(n-k) e^{j(2\pi f_c nT + \theta(n))} \quad (6k)$$

$$B^{(11)}_{k1,k2,k3}(n+1) = B^{(11)}_{k1,k2,k3}(n) + \gamma_{11} \epsilon(n)^* \hat{A}(n-k_1)\hat{A}(n-k_2)\hat{A}(n-k_3)^* e^{j(2\pi f_c nT + \hat{\theta}(n))} \quad (6l)$$

$$B^{(12)}_{k1,k2,k3}(n+1) = B^{(12)}_{k1,k2,k3}(n) + \gamma_{12} \epsilon(n)^* \hat{A}(n-k_1)^* \hat{A}(n-k_2)^* \hat{A}(n-k_3) e^{-j(2\pi f_c nT + \hat{\theta}(n))} \quad (6m)$$

$$B^{(2)}_{k1,k2}(n+1) = B^{(2)}_{k1,k2}(n) + \gamma_2 \epsilon(n)^* \hat{A}(n-k_1) \cdot \hat{A}(n-k_2) e^{j(4\pi f_c nT + \hat{\theta}(n))} \quad (6n)$$

$$B^{(3)}_{k1,k2,k3}(n+1) = B^{(3)}_{k1,k2,k3}(n) + \gamma_3 \epsilon(n)^* \hat{A}(n-k_1)\hat{A}(n-k_2)\hat{A}(n-k_3) e^{j(6\pi f_c nT + \hat{\theta}(n))}. \quad (6o)$$

The set of eqs. (4) through (6) defines the structure of the nonlinear QAM receiver that has been simulated. The α, β, and γ parameters are positive constants, chosen to ensure reasonably fast convergence and stability in the presence of noise. To enable compensation of rapidly varying phase jitter, the phase tracking constant α was set to the relatively large value of 0.4. The other constants chosen were:

$$\beta_1 = \gamma_1 = 0.001, \beta_0 = \beta_2 = \gamma_0 = \gamma_2 = 0.75 \times 10^{-5},$$
$$\beta_{11} = \beta_{12} = \beta_3 = \gamma_{11} = \gamma_{12} = \gamma_3 = 10^{-6}.$$

A judicious choice must be made for the range of coefficient indices k_1, k_2, and k_3 in the nonlinear terms making up $Y_{NL}(n)$, if the total number of $\{W\}$ and $\{B\}$ coefficients is to be reasonable, say on the order of 100. Obviously, the best choice of indices for a fixed number of taps depends on the channel. Trial and error (by no means exhaustive) of various sets of indices used in simulations on several voiceband channels led to the choice of terms shown in Table I. There are 73 "forward" tap coefficients $\{W\}$, of which 22 are linear, and 61 "feedback" tap coefficients $\{B\}$, of which 10 are linear. Note that the nonlinear forward tap indices are confined to the range $-1 \leq k \leq 1$ and the nonlinear feedback tap indices have been confined to the range $1 \leq k \leq 3$.

VI. THE SIMULATIONS

The nonlinear QAM receiver structure described in the previous section was simulated on an IBM 360 computer to process recorded 9600-bps QAM data signals that had been received from 17 voiceband telephone channels. The simulation effort was an extension of that described for linear and decision feedback QAM receivers in Refs. 3 and 9, respectively. The set of recorded QAM signals was the same, permitting the performance of all three receiver types to be compared under identical con-

Table I — Index terms used in voiceband simulations

Terms	Indices k_1	k_2	k_3	Terms	Indices k_1	k_2	k_3
$W^{(0)}_{k1,k2}$	−1	−1		$B^{(0)}_{k1,k2}$	1	1	
	0	0			2	2	
	1	1			3	3	
	−1	0			2	1	
	0	−1			1	2	
	0	1			3	2	
	1	0			2	3	
	−1	1			3	1	
	1	−1			1	3	
$W^{(1)}_{k1}$ (Linear) terms −12 to 9 inclusive				$B^{(1)}_{k1}$ (Linear) terms 1 to 10 inclusive			
$W^{(11)}_{k1,k2,k3}$				$B^{(11)}_{k1,k2,k3}$ and			
and $W^{(12)}_{k1,k2,k3}$	−1	−1	−1	$B^{(12)}_{k1,k2,k3}$	1	1	1
	0	0	0		2	2	2
	1	1	1		3	3	3
	−1	−1	0		1	1	2
	−1	0	−1		1	2	1
	0	0	−1		2	2	1
	0	−1	0		2	1	2
	0	0	1		2	2	3
	0	1	0		2	3	2
	1	1	0		3	3	2
	1	0	1		3	2	3
	−1	0	1		1	2	3
	0	1	−1		2	3	1
	−1	1	0		1	3	2
$W^{(2)}_{k1,k2}$	−1	−1		$B^{(2)}_{k1,k2}$	1	1	
	−1	0			2	2	
	0	0			3	3	
	1	1			1	2	
	0	1			2	3	
	−1	1			1	3	
$W^{(3)}_{k1,k2,k3}$	−1	−1	−1	$B^{(3)}_{k1,k2,k3}$	1	1	1
	0	0	0		2	2	2
	−1	−1	0		1	1	2
	0	0	−1		2	2	1
	1	1	1		3	3	3
	0	0	1		2	2	3
	1	1	0		3	3	2
	−1	0	1		2	3	3

ditions. The set of 17 channels could be described as "worse than average." Every channel had at least one impairment equal to or worse than the 90-percent point on the nationwide toll connection survey.[2]

The transmitted QAM signals had been generated digitally, with two pseudorandom four-level information symbol streams in quadrature, each repeating after 256 symbols. Each quadrature pair of symbols

therefore conveyed four information bits and the symbol rate was 2400 bauds, making a total bit rate of 9600 bps. The carrier frequency f_c was 1650 Hz, and the double-sideband baseband pulse signal had 12 percent roll-off.

The received signals that were recorded in digital form (12-bit samples, 24-kHz sampling rate) were received from a variety of real and analog-simulated voiceband telephone channels in tandem with an actual 50-km, C2-conditioned, N2-carrier voiceband channel.

As in the simulation of the linear and decision feedback receivers, the adaptive passband signal processors [defined by the set of eqs. (4) and Table I] were preceded by a pair of fixed digital filters that split the incoming signal into in-phase and quadrature components. Each was sampled at time instants $t = \tau + nT$ ($n = 0,1,2,\cdots$). Each simulation was actually of five separate receivers in parallel, with sampling epochs $\tau =$ 0, 0.2T, 0.4T, 0.6T, and 0.8T. The results reported in this paper are in each case for the timing epoch which yielded the best performance. As noted previously in Ref. 9, the decision feedback structure generally produced a relatively small performance spread between the best and the worst timing epochs. The receiver's decisions $\hat{A}(n)$ were formed by quantizing each equalized demodulated output, in-phase or quadrature, into one of the four possible levels $\pm 1, \pm 3$.

Before tabulating the simulation results, we mention some qualitative observations. In the interest of reducing the large numbers of nonlinear coefficients, it would have been desirable that only a few of the observed coefficients be large enough to be significant for all the channels. Unfortunately, this was not the case; no pattern was discernible common to all channels of a significant subset of coefficients; typically, the nonlinear component $Y_{NL}(n)$ in the receiver's output consisted of a large number of small terms, rather than a small number of relatively large terms plus insignificant terms.

Another qualitative observation was that the best values for the adaptation parameters for the nonlinear coefficients were so small that convergence of the nonlinear tap coefficients required at least 2000 symbol intervals, much slower than the convergence rate of the linear coefficients. This is attributed to the high correlation among many of the nonlinear terms. For example, the term $|A_{k_1}|^2 A_{k_2}$ is positively correlated with the linear term A_{k_2}, since $|A_{k_1}|^2$ takes only one of the three possible positive values 2, 10, or 18. Under such circumstances, the $\mathcal{A}$ matrix which describes the correlations among all the terms is expected to have a rather large eigenvalue spread, necessitating small adaptation constants and slow convergence.[13]

During each run, after an initial training period of 2000 symbol intervals to allow the coefficients to converge to nearly stationary values, the simulated receivers switched to a decision-directed mode in which

their decisions $\hat{A}(n)$, right or wrong, were used in the adaptation and decision feedback operations. Since the true transmitted information stream $\{A(n)\}$ was known, the performance was measured by observing the number of decision errors made during 7000 symbol intervals (or 28,000 bits). The empirical probability of the sampled analog error $Y(n) - A(n)$ was also measured, and if no errors were observed during a run, the error probability could be roughly estimated by extrapolating the tail of this distribution, using a computer subroutine by S. B. Weinstein.[14] The tabulated error probability, p_e, is the probability that a four-level symbol is in error; i.e., it is roughly twice the bit error rate. Another tabulated measure of performance was the *output SNR*, defined by

$$\text{output SNR} = \frac{\langle |A(n)|^2 \rangle}{\langle |E(n)|^2 \rangle}$$

where "$\langle \quad \rangle$" denotes the time average.

VII. QUANTITATIVE RESULTS

The simulation results for the NL receiver are tabulated in Table II along with the corresponding results taken from Ref. 9 for the LE and DFE receivers. For each channel, Table II lists the measured impairments and the error probabilities (either observed or extrapolated) for the LE, DFE, and NL receivers. The quantity in parentheses below each error probability is the output SNR in decibels. Error rates below 10^{-5} were extrapolated; in some cases in which the tail of the empirical probability distribution of the quadrature components of $E(n)$ was markedly non-Gaussian, the extrapolation yielded limited accuracy. Figure 4 illustrates the nonlinear compensation for channel 14, which had unusually severe second-harmonic distortion. Figure 4 is plotted on a "probability scale;" i.e., a Gaussian error distribution function would plot as a straight line on it. The distribution function for the linear receiver has distorted tails, indicating the presence of residual nonlinear distortion. However, the curve is nearly straight for the NL receiver, indicating that nonlinear distortion components have been substantially removed.

Comparison of error rates for the three receivers on all the channels is displayed more dramatically by the bar graph of Fig. 5. In all cases, the performance of the NL receiver surpassed that of the other two receivers. (Note that measurable nonlinear distortion was observed on all the channels.) In most cases, the NL receiver afforded a greater improvement in error rate over the DFE receiver than did the DFE receiver over the LE receiver. This is a very significant point. It indicates that if 9600-bps voiceband modems are to be improved by more sophisticated signal processing at the receiver, it is more fruitful to attempt to overcome nonlinear distortion than to concentrate on more sophisticated receiver structures, optimal for linear channel models.

Table II — Experimental comparison of LE, DFE and NL receivers
(facility in tandem with Holmdel-Murray Hill N2-carrier line)

		No.5 none	No. 6 Private N Carrier to White Plains	No. 7 Line Simulator	No. 8 Private T1 Carrier to Newark	No. 9 Line Simulator	No. 10 Line Simulator	No. 11 Line Simulator	No. 12 Line Simulator	No. 13 Line Simulator
Measured Impairments	Slope (dB)	0	2	9	4.4	−2	3	11*	11*	0
	Signal-to-noise ratio (dB)	29.0*	22.5†	31	27*	30	35	24.4*	33	34
	Second harmonic (dB)	33.5	28*	33	35	25†	32.1*	28.6*	33.8	34.4
	Third harmonic (dB)	44	31*	22.5†	40	33	47	36.4	49	30.7*
	Phase jitter (peak-to-peak)	<3°	<3°	<3°	<3°	14°† (120 Hz)	17°† (50 Hz)	<3°	<3°	<3°
Error rates (output SNR)	Linear equalization	1×10^{-8} (28.0 dB)	2×10^{-6} (22.4 dB)	6×10^{-3} (17.9 dB)	1×10^{-5} (20.8 dB)	7×10^{-3} (15.0 dB)	8×10^{-6} (23.9 dB)	1×10^{-4} (18.6 dB)	3×10^{-7} (22.6 dB)	3×10^{-6} (23.8 dB)
	Decision feedback equalization	2×10^{-9} (27.8 dB)	4×10^{-6} (22.8 dB)	3×10^{-3} (19.5 dB)	4×10^{-6} (21.2 dB)	9×10^{-3} (15.0 dB)	5×10^{-7} (23.8 dB)	7×10^{-5} (20.2 dB)	7×10^{-10} (24.6 dB)	3×10^{-8} (23.9 dB)
	Nonlinearity equalization	2×10^{-11} (29.4 dB)	1×10^{-6} (24.9 dB)	1×10^{-5} (23.7 dB)	7×10^{-7} (22.3 dB)	2×10^{-4} (18.2 dB)	3×10^{-8} (24.7 dB)	3×10^{-7} (21.7 dB)	3×10^{-12} (28.6 dB)	1×10^{-12} (27.5 dB)

Table II (cont)

		No. 14 Line Simulator	No. 15 Line Simulator	No. 16 Line Simulator	No. 17 Line Simulator	No. 18 DDD Loopback to Dallas	No. 19 (Private T1 Carrier to Newark	No. 20 Private T1 Carrier to Newark	No. 21 Private T1 Carrier to Newark
Measured Impairments	Slope (dB)	0	0	12†	11.1*	7.8	13†	6	8
	Signal-to-noise ratio (dB)	31	31	23†	29*	29*	28*	24.8*	23.2†
	Second harmonic (dB)	20.6†	27.2*	25.2†	32.2*	36.4	31.8*	24.4†	24.6†
	Third harmonic (dB)	49	32*	30.3†	34.7*	32.2*	37	32.6*	28.6†
	Phase jitter (peak-to-peak)	<3°	<3°	15°† (120 Hz)	10°† (120 Hz)	6°	<3°	<3°	<3°
Error Rates (output SNR)	Linear equalization	2×10^{-4} (19.4 dB)	1×10^{-6} (24.5 dB)	1.7×10^{-2} (14.0 dB)	3×10^{-3} (17.4 dB)	1×10^{-3} (18.5 dB)	2.1×10^{-3} (18.3 dB)	5×10^{-4} (18.4 dB)	1.6×10^{-3} (17.8 dB)
	Decision feedback equalization	8×10^{-5} (19.5 dB)	5×10^{-7} (24.5 dB)	3×10^{-2} (14.1 dB)	2×10^{-3} (18.6 dB)	9×10^{-4} (18.3 dB)	2×10^{-3} (19.9 dB)	3×10^{-4} (18.7 dB)	3×10^{-3} (18.0 dB)
	Nonlinearity equalization	2×10^{-9} (26.8 dB)	1×10^{-10} (28.4 dB)	1×10^{-2} (15.6 dB)	3×10^{-5} (20.8 dB)	5×10^{-5} (20.4 dB)	5×10^{-5} (21.6 dB)	3×10^{-5} (20.6 dB)	7×10^{-5} (19.6 dB)

* Indicates worse than 90-percent point in the nationwide toll connection survey.
† Indicates worse than "worst case" 3002 channel impairment limit.

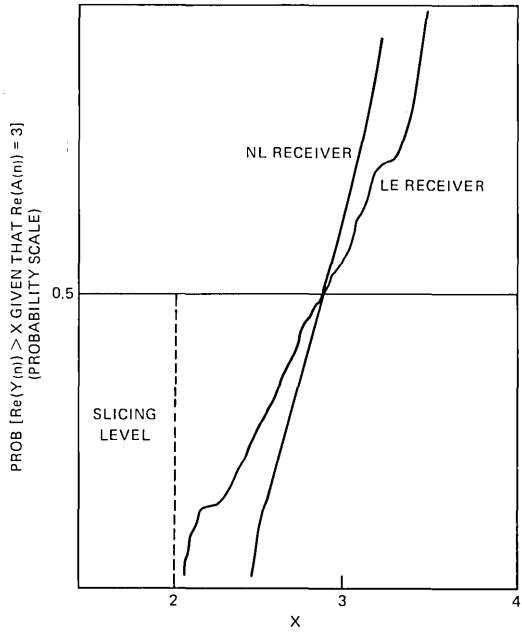


Fig. 4—Comparison of distribution functions of the receiver output $Y(n)$ for the linear and nonlinear receivers (data from channel 14).

Note in Fig. 5 that, for some of the channels, the nonlinearity equalization reduced the error rate by two or three orders of magnitude. However, on other channels, such as 9 and 16* which had most of their impairments in the "severe" category, the error rate was high and the NL receiver afforded very little improvement.

An interesting statistic that can be gleaned from Fig. 5 concerns the ability of the NL receiver to increase the number of channels which yield error rates below a specified maximum. For example, 15 of the 17 channels yield an error rate of better than 10^{-4} with the NL receiver, but only 8 of 17 meet this error rate standard with the LE receiver. For a maximum error rate of 10^{-5}, the number of channels is 10 with the NL receiver and 7 with the LE receiver. For a maximum error rate of 10^{-6}, the numbers of channels are 9 and 3 with the NL and LE receivers, respectively.

The price paid for the better performance of the NL receiver is, of course, its increased complexity, measured by the number of terms comprising $Y_{NL}(n)$ in eq. (4) and its slower convergence. The effect of reducing the number of terms, and therefore the complexity, is treated in the next section.

* Channel 16's impairments, produced by a line simulator, were all "worst case" values.

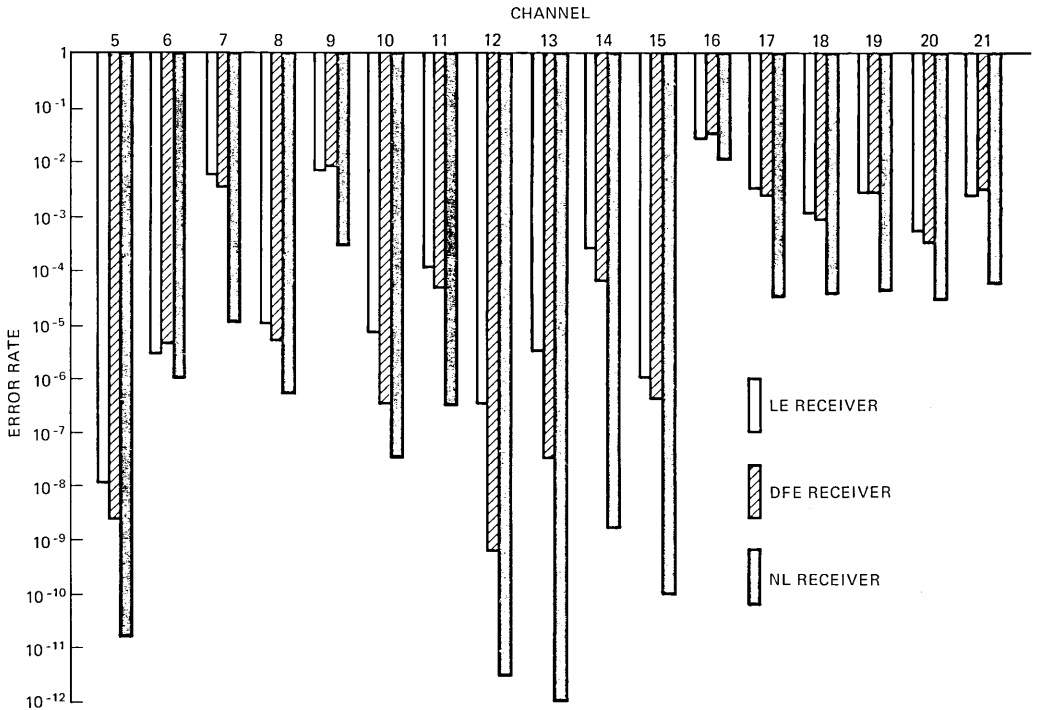


Fig. 5—Comparison of error rates for the three receivers.

VIII. MODIFICATIONS OF THE NONLINEAR RECEIVER STRUCTURE

8.1 Reductions of the number of nonlinear tap coefficients

(*i*) The tap coefficients $\{W^{(12)}_{k1,k2,k3}\}$ and $\{B^{(12)}_{k1,k2,k3}\}$ were set to zero, reducing the total number of nonlinear forward and feedback taps to 37 each. The measured output SNRs for most of the channels were slightly less than those for the full complement of 51 forward and 51 feedback taps, as illustrated in Table III.

(*ii*) A different set of 100 nonlinear terms was created by eliminating all cross-product terms and extending the time span covered by the forward and feedback terms to 10 symbol intervals. Thus, the forward tap coefficients consisted of $\{W^{(0)}_{k,k}\}$, $\{W^{(11)}_{k,k,k}\}$, $\{W^{(12)}_{k,k,k}\}$, $\{W^{(12)}_{k,k}\}$, and $\{W^{(13)}_{k,k,k}\}$, where $-5 \leq k \leq 4$, and the feedback terms consisted of $\{B^{(0)}_{k,k}\}$, $\{B^{(11)}_{k,k,k}\}$, $\{B^{(12)}_{k,k,k}\}$, $\{B^{(2)}_{k,k}\}$, and $\{B^{(3)}_{k,k,k}\}$, where $1 \leq k \leq 10$. Some resulting output SNRs are tabulated in part (*iii*) following.

(*iii*) A smaller set of nonlinear taps was created by taking a subset of 46 of the original set of 102 nonlinear taps. The resulting output SNRs for several channels are shown in Table IV, along with the corresponding set of SNRs from the original NL receiver structure with 102 nonlinear taps and also from the receiver with 100 nonlinear taps, described in item (*ii*), above.

The results of items (*i*), (*ii*), and (*iii*), compared with the original results using the NL receiver with 102 nonlinear tap coefficients indicate that a large number of nonlinear correction terms is necessary to yield substantial performance improvement. Undoubtedly, still better performance would have been attained by using more than 102 nonlinear taps. The results of item (*ii*) also showed that elimination of the cross-product terms degraded performance, even though the remaining nonlinear terms encompassed a longer time span.

Table III — Output SNR (dB) for nonlinear receivers

Channel	102 nonlinear taps	74 nonlinear taps
5	29.4	29.4
6	24.9	23.8
7	23.7	22.7
8	22.3	22.1
9	18.2	17.5
10	24.7	24.5
11	21.7	21.5
12	28.6	27.5

Table IV — Output SNR (dB) for nonlinear receivers

Channel	Original (102 Taps)	(*ii*) 100 Taps	(*iii*) 46 Taps
9	18.2	17.7	17.5
13	27.5	25.8	25.7
14	26.8	23.7	23.0
15	28.4	25.6	26.5

(*iv*) The number of nonlinear taps was also reduced to 46 by eliminating all coefficients $\{W^{(2)}_{k1,k2}\}$, $\{B^{(2)}_{k1,k2}\}$, $\{W^{(3)}_{k1,k2,k3}\}$, and $\{B^{(3)}_{k1,k2,k3}\}$. The resulting output SNR on channel 14 was only 21.3 dB, as compared to 24.8 dB for 102 nonlinear taps. Thus, it appears that at least the last four sets of coefficients (associated with second and third harmonics of the carrier frequency) are significant and should be retained.

8.2 A variation in the receiver structure tested for channel 20

The forward nonlinear tap coefficients weight various quadratic and cubic products of the sampled received signals. One might speculate that if *linear* distortion were removed from the received samples before their nonlinear processing, the nonlinear distortion remaining in the output might be further reduced. Accordingly, we simulated an NL receiver structure which was the same as that shown in Fig. 4 except that there are no linear feedback taps and the input to the forward nonlinear taps comes from the output of the linear forward taps instead of directly from the phase splitter. Since the adaptive linear forward taps, constituting the passband equilizer, are in tandem with the adaptive nonlinear taps in this structure, the mean squared error is not a convex function of the nonlinear tap coefficients, and hence the question of convergence is more complicated. Nevertheless, this structure was simulated on channel 20. The resulting output SNR was 20.0 dB compared to the 20.6 dB obtained from the original receiver structure. Thus, prior linear equalization did not appear preferable.

IX. CONCLUSIONS

The simulations have demonstrated that nonlinearity-equalizing QAM receivers can provide substantially better performance than can conventional linear or decision feedback equalization receivers over a variety of voiceband telephone channels. This encouraging result may stimulate further research aimed at finding less complicated receiver structures for overcoming channel nonlinearities.

The number of nonlinear terms that can be considered for inclusion in the NL receiver's analog output $Y(n)$ is potentially enormous. For example, the number of different terms $R(k_1)R(k_2)R(k_3)^*$ for all indices k_1, k_2 and k_3 between $-N$ and $+N$ is $(2N+1)^2(N+1)$, which is much more than $(2N+1)$, the corresponding number of linear terms $\{R(k)\}$ in that range of indices. The simulation results indicated that inclusion of a large number of nonlinear terms, including "cross-product" terms for which $k_1 \neq k_2 \neq k_3$, may be necessary. Reductions in the number of terms and a variation of the NL receiver's structure, in which adaptive linear processing preceded nonlinear processing, resulted in worsened performance.

Perhaps the major conclusion to be drawn concerns means for im-

proving the reliability of high-speed data transmission over the population of voiceband telephone channels. The simulations reported in Ref. 9 showed that decision feedback equalization, which is known theoretically to be superior to linear equalization in overcoming severe linear distortion, only moderately bettered the error rate obtained with linear equalization, especially on voiceband channels meeting C2 conditioning standards. However, the results summarized by Fig. 5 indicated that there is more to be gained by mitigating nonlinear distortion than in using more elaborate methods (beyond linear or decision feedback equalization) of mitigating linear distortion.

APPENDIX

In this appendix, we derive the form of the analytic signal that emerges from the summed filtered outputs of the quadratic and cubic nonlinearities. The real and imaginary parts of this analytic signal will then be the in-phase and quadrature components, respectively, of the nonlinearly distorted received QAM signal. The following theorems, proven in Ref. 11, will be required:

Theorem 1: Given real waveforms $u(t)$ and $v(t)$, defined on $-\infty < t < \infty$ with respective Hilbert transforms $\check{u}(t)$ and $\check{v}(t)$, the convolution

$$w(t) = \int_{-\infty}^{\infty} v(\tau)u(t-\tau)d\tau \tag{7}$$

has Hilbert transform

$$\check{w}(t) = \int_{-\infty}^{\infty} \check{v}(\tau)u(t-\tau)d\tau = \int_{-\infty}^{\infty} v(\tau)\check{u}(t-\tau)d\tau. \tag{8}$$

Thus, if $v(t)$ is the input to a filter whose impulse response is $u(t)$, the analytic output signal is

$$\begin{aligned} w(t) + j\check{w}(t) &= \int_{-\infty}^{\infty} (v(\tau) + j\check{v}(\tau))u(t-\tau)d\tau \\ &= \int_{-\infty}^{\infty} v(\tau)(u(t-\tau) + j\check{u}(t-\tau))d\tau. \end{aligned} \tag{9}$$

Theorem 2: The analytic signal resulting from the convolution can also be expressed as

$$w(t) + j\check{w}(t) = \frac{1}{2}\int_{-\infty}^{\infty} (v(\tau) + j\check{v}(\tau))(u(t-\tau) + j\check{u}(t-\tau))d\tau. \tag{10}$$

Now we consider an analytic signal of the form

$$X(t) = e^{j2\pi f_c t}\sum_n A(n)F(t-nT), \tag{11}$$

as in expression (2) of the text. The squaring and cubing elements in Fig. 2 operate on $x(t)$, the real part of $X(t)$. The response of the squaring element to $\mathrm{Re}(X(t))$ can be written

$$x(t)^2 = \frac{1}{2}\mathrm{Re}\left[e^{j4\pi f_c t}\sum_{n_1,n_2} A(n_1)A(n_2)F(t-n_1T)F(t-n_2T)\right] + \frac{1}{2}\left[\sum_{n_1,n_2} A(n_1)A(n_2)^*F(t-n_1T)F(t-n_2T)^*\right]. \quad (12)$$

Of the complex expressions in square brackets in (12), the first is complex and analytic, since it is the square of an analytic signal (its spectrum is nonzero only for positive frequencies). Thus, from Theorem 2, the analytic signal that results from passing the first part of expression (12) through a passband filter 2 is of the form

$$U_2(t) = e^{j4\pi f_c t}\sum_{n_1}\sum_{n_2} A(n_1)A(n_2)G_2(t-n_1T,t-n_2T), \quad (13)$$

where $G_2(t-n_1T,t-n_2T)e^{j4\pi f_c t}$ is a complex analytic waveform, whose spectrum has been limited by filter 2 to $0 < f < 2f_c$. The second term in (12) is baseband, real, and not analytic.† However, from Theorem 1, the analytic signal resulting from passing the second term through filter 2 has the form

$$U_0(t) = \sum_{n_1,n_2} A(n_1)A(n_2)^*G_0(t-n_1T,t-n_2T), \quad (14)$$

where $G_0(t-n_1T,t-n_2T)$ is an analytic waveform, whose spectrum is confined to $0 < f < 2f_c$.

The cubic nonlinear terms are handled similarly. The cube of the input signal $\mathrm{Re}(X(t))$ can be written

$$x(t)^3 = \frac{1}{4}\mathrm{Re}\left[e^{j6\pi f_c t}\sum_{n_1,n_2,n_3} A(n_1)A(n_2)A(n_3)F(t-n_1T)\cdot F(t-n_2T)F(t-n_3T)\right] + \frac{3}{8}e^{j2\pi f_c t}\sum_{n_1,n_2,n_3} A(n_1)A(n_2)A(n_3)^*F(t-n_1T)F(t-n_2T)\cdot F(t-n_3T)^* + \frac{3}{8}e^{-j2\pi f_c t}\sum_{n_1,n_2,n_3} A(n_1)^*A(n_2)^*A(n_3)\cdot F(t-n_1T)^*F(t-n_2T)^*F(n-n_3T). \quad (15)$$

The first term in square brackets (15) is analytic, being the cube of an analytic signal. The other two terms in (15) are not analytic, since their

† The ranges of the indices n_1 and n_2 in (12) and (13) are assumed to be the same.

Fourier transforms are not necessarily zero for negative frequencies. The analytic signal resulting from passing $x(t)^3$ through bandpass filter 3 can be written by applying Theorem 2 to the first term of (15) and Theorem 1 to the second and third terms. The resulting analytic signal is the sum of three analytic signals, $U_3(t)$, $U_{11}(t)$, and $U_{12}(t)$, which have the following forms:

$$U_3(t) = e^{j6\pi f_c t} \sum_{n_1,n_2,n_3} A(n_1)A(n_2)A(n_3) \cdot G_3(t - n_1T, t - n_2T, t - n_3T). \quad (16)$$

$$U_{11}(t) = e^{j2\pi f_c t} \sum_{n_1,n_2,n_3} A(n_1)A(n_2)A(n_3)^* \cdot G_{11}(t - n_1T, t - n_2T, t - n_3T). \quad (17)$$

$$U_{12}(t) = e^{-j2\pi f_c t} \sum_{n_1,n_2,n_3} A(n_1)^*A(n_2)^*A(n_3) \cdot G_{12}(t - n_1T, t - n_2T, t - n_3T). \quad (18)$$

The $G(\ )$ signals are complex, and the spectra of the analytic signals $U_3(t)$, $U_{11}(t)$, and $U_{12}(t)$ are all confined to the range $0 < f < 2f_c$ by bandpass filter 3.

REFERENCES

1. R. W. Lucky, "Modulation and Detection for Data Transmission on the Telephone Channel," in *New Directions in Signal Processing in Communication and Control*, ed. by J. K. Skwirzynski, Leyden: Noordhoff, 1975, pp. 321–327.
2. F. P. Duffy and T. W. Thatcher, Jr., "Analog Transmission Performance on the Switched Telecommunications Network," B.S.T.J., *50*, No. 4 (April 1971), pp. 1311–1347.
3. R. R. Anderson and D. D. Falconer, "Modem Evaluation on Real Channels Using Computer Simulation," Proc. National Telecomm. Conf., Dec. 1974, San Diego, pp. 877–883.
4. S. Benedetto, E. Biglieri, and R. Daffara, "Performance of Multilevel Baseband Digital Systems in a Nonlinear Environment," IEEE Trans. Comm., *COM-24*, No. 10 (October 1976), pp. 1166–1175.
5. W. J. Lawless and M. Schwartz, "Binary Signaling over Channels Containing Quadratic Nonlinearities," IEEE Trans. Comm., *COM-22*, No. 2 (March 1974), pp. 288–298.
6. N. Wiener, *Nonlinear Problems in Random Theory*, Cambridge, Mass.: M.I.T. Press and J. Wiley, 1958.
7. T. Arbuckle, "Nonlinear Equalization System Including Self- and Cross-Multiplication of Sampled Signals," U.S. Patent 3,600,681, Aug. 17, 1971.
8. E. J. Thomas, "Some Considerations on the Application of the Volterra Representation of Nonlinear Networks to Adaptive Echo Cancellers," B.S.T.J., *50*, No. 8 (October 1971), pp. 2797–2805.
9. D. D. Falconer, "Application of Passband Decision Feedback Equalization in Two-Dimensional Data Communications Systems," IEEE Trans. Comm., *COM-24*, No. 10 (October 1976), pp. 1159–1166.
10. D. D. Falconer, "Jointly Adaptive Equalization and Carrier Recovery in Two Dimensional Digital Communication Systems," B.S.T.J., *55*, No. 3 (March 1976), pp. 317–334.
11. J. Dungundji, "Envelopes and Pre-Envelopes of Real Waveforms," IRE Trans. on Info. Theory, Sept. 1957, pp. 53–57.
12. R. D. Gitlin, E. Y. Ho, and J. E. Mazo, "Passband Equalization for Differentially Phase Modulated Data Signals," B.S.T.J., *52*, No. 2 (February 1973), pp. 219–238.

13. A. Gersho, "Adaptive Equalization of Highly Dispersive Channels for Data Transmission," B.S.T.J., *48,* No. 1 (January 1969), pp. 55–70.
14. S. B. Weinstein, "Theory and Applications of Some Classical and Generalized Asymptotic Distributions of Extreme Values," IEEE Trans. Inform. Theory, *IT-19,* No. 2 (March 1973), pp. 148–154.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 7, September 1978
Printed in U.S.A.

Spectral Sharing in Hybrid Spot and Area Coverage Satellite Systems via Channel Coding Techniques

By A. S. ACAMPORA

(Manuscript received December 1, 1977)

Multiple spot-beam switching satellites employing frequency reuse are considered, and a method for incorporating an area coverage beam to provide service to those regions not covered by the footprint of any spot beam is proposed here. The method consists of employing a convolutional code for the area beam transmission to enable sharing of a common spectral band among the spot and area beams on a noninterfering basis and with no sacrifice in the capacity of the spot beams. A maximum-likelihood algorithm for this purpose is derived, and bounds on the bit error rate performance of all beams are found. Results show that excessive performance degradation arising from cochannel interference is limited to a thin annular ring surrounding each spot beam.

I. INTRODUCTION

Multiple spot beam communication satellites offer the potential for greatly increasing the traffic handling capability relative to wide-area coverage systems, since the allocated spectral band can be reused in the various spot beams.[1,2] A high-level block diagram of the satellite transponders for such a system might appear as shown in Fig. 1. Here, the various service regions are interconnected via an on-board switching matrix operating in the time-division mode, and digital modulation techniques consistent with time-division multiple access (TDMA) are employed.

As previously noted,[3] such a system suffers a serious drawback in that a large blackout region, serviceable by none of the spot beams, is created. The situation is depicted in Fig. 2, which shows the radiation footprints of a hypothetical 11-beam private line system serving the large population regions in the United States. Although most of the traffic load for

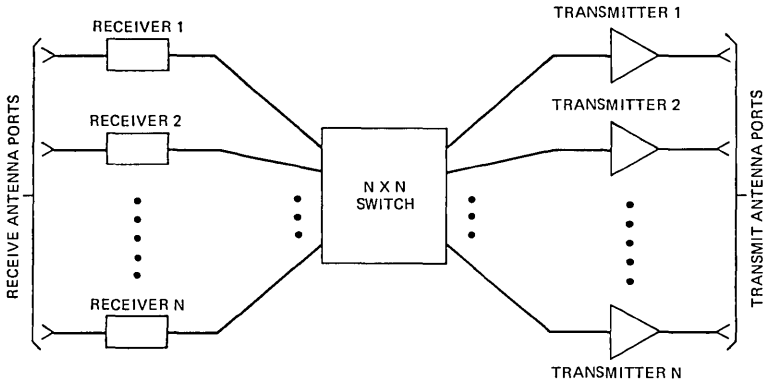


Fig. 1—Satellite transponder.

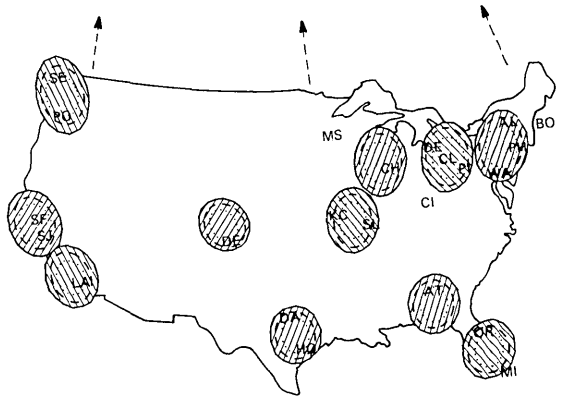


Fig. 2—Footprints of a hypothetical 11-beam system showing −1, −2, and −3 dB contours. Both polarizations are employed.

such an offering would be adequately served by the 11 high-capacity spot beams, it is nonetheless desirable to provide service to the outlying areas.

Among the various techniques proposed in Ref. 3 for coping with this blackout problem, the method of deploying a channel coded area coverage beam, in addition to the various uncoded spot beams, appears most attractive in that the blackout region is reduced to a thin annular ring surrounding each spot beam. This method offers the additional advantage of reducing the required radiated power for the area coverage beam, an important consideration since the gain of the area beam antenna port might be 20 dB lower than that of a spot beam port. In this paper, we review the principles involved in this approach and derive bounds on the resulting bit error rate performance of both the spot and area beams.

In Section II, we discuss the problems associated with sharing a

common spectral band between area coverage and spot coverage satellite beams. Section III is devoted to the derivation of a detection algorithm for such a hybrid system in which convolutional coding is used to alleviate the effects of cochannel interference, and bit error rate bounds are found. In Section IV, these results are applied in a typical communication satellite scenario.

II. PROBLEM DEFINITION

Consider a satellite system consisting of M spot-beam transponders serving M geographically separated, high-traffic demand areas on a noninterfering basis. The allocated spectral band is totally reused in the M spot beams. We wish to deploy an area coverage beam, in addition to the M spot beams, to provide service to the low traffic demand outlying regions serviced by none of the spot beams. The total traffic demand to all outlying regions might be of the same order of magnitude as the demand for one spot beam. Service to the outlying regions must be provided on a noninterfering basis and with no sacrifice to the capacity of the various spot beams. We assume that the spot beams require use of both electromagnetic polarizations to minimize mutual interference among themselves.

Four types of interference are readily identified:

(*i*) *Down-link:* The area coverage radiation is detectable at every spot beam receiving terminal and can thereby interfere with reception of the desired signal at those ground stations.

(*ii*) *Down-link:* The spot beam footprints might typically be useful out to their −3 dB radiation contours. Area-coverage receiving terminals located at the −3 dB through the −20 dB contours of any spot beam thereby suffers interference from that spot beam.

(*iii*) *Up-link:* All up-link transmissions from spot-beam earth terminals are detectable at the antenna port of the area coverage beam and thereby interfere with reception of the area coverage up-link transmission.

(*iv*) *Up-link:* Transmission from an area coverage ground station located between the −3 dB and the −20 dB contour of a spot-beam antenna pattern could interfere with that spot beam's up-link transmission.

Thus, the inclusion of an area beam might make the original spot beams totally unusable. To eliminate these interference problems, one might split the allocated spectral band into two components; one segment would be dedicated to the area coverage transponder and the second segment would be reused among the various spot beams. If this is done, the system designer must choose one of two options:

(*i*) Reduce the throughput of the spot beam transponders by that fraction of the satellite band dedicated to the area coverage beam.

(*ii*) Maintain the original throughput of the spot beam transponders while increasing the effective radiated power on both the up-link and the down-link to overcome the degradation caused by excessive band-limiting.

Option (*i*) results in a sizable decrease in the overall capacity of the satellite. Consider a 10-spot-beam system with each beam occupying the entire spectral band. The normalized throughput of such a system is defined to be 10 units. Suppose that two area coverage beams are added to the system (one using each polarization) and that one-half the band is reserved for the spot beams. Then, under option (*i*), the normalized throughput is reduced to $\frac{1}{2} \times 10 + \frac{1}{2} \times 2 = 6$, and the overall system throughput is reduced by 40 percent. For the same fractional split of the total bandwidth, option (*ii*) could incur a power penalty in excess of 6 dB for a 4ϕ-CPSK (coherent phase shift key) system originally operating at a modest BT (bandwidth-time) product of 1.3. Such a penalty might be acceptable on the up-link, but would typically be unacceptable on the down-link since space platform power is a limited resource.

Thus, to provide service to the outlying area at no sacrifice in either the throughput of the spot beams or in the required spot beam effective isotopic radiated power (e.i.r.p.), one might consider splitting the band, as described above, to eliminate up-link interference. Up-link digits would be regenerated, switched, and reformatted into the appropriate down-link port. A suitable scheme must then be sought to accommodate the down-link.

Channel coding techniques will be investigated as a possibility. The motivation for such an approach is twofold. First, and most important, coding can provide for effective immunity against cochannel interference. Second, we note that because of the difference of about 20 dB between the antenna gains of the area and spot-beam coverage antenna ports, a system would require 20 dB more power for the area coverage port than for a spot-beam port to achieve the same bit error rate performance. Through use of coding, we can effect a considerable reduction in the required power for the global beam.

The scenario envisioned, shown in Fig. 3, would employ uncoded transmission for the spot-beam messages and rate $r = \frac{1}{2}$ convolutionally encoded transmission for the area beam port. The throughput of the area beam port would be one-half that of the spot-beam port, implying that the down-link channel symbol rate for all beams are the same. In addition, since on-board regeneration is employed, all down-link channel symbols can be time-aligned. We will consider 4ϕ-CPSK modulation and explore in detail the situation where the in-phase and quadrature rails of both the area and spot coverage beams are modulated separately and where there is no crossrail coupling. Thus, we can consider baseband performance. The algorithms and other results to be presented are

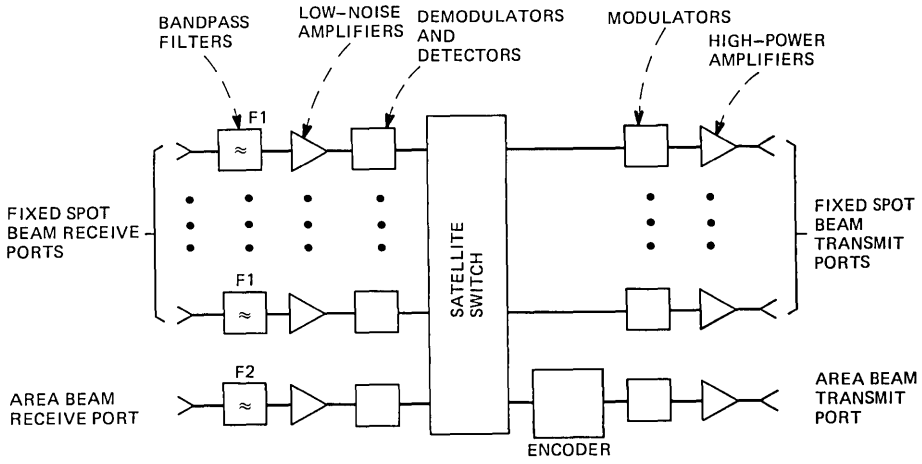


Fig. 3—Regenerative transponder for a hybrid spot-area coverage satellite employing channel coding. The allocated band is split between the spot and area beams on the uplink. The band is totally reused for all the downlinks.

readily generalized to the situation where there is a known, fixed carrier phase shift between the spot and area transmissions.

III. BIT ERROR RATE PERFORMANCE

We now investigate the bit error rate performance of both the uncoded spot beam message and the encoded global beam message in the interference-prone region surrounding one of the spot beams. We need to consider the presence of only one such spot beam since, in the footprint area of that beam, interference from the remaining beams is negligible. At a particular ground station, after coherent demodulation, we observe the following received baseband process:

$$R(t) = \sqrt{E_1} \sum_k b_k h(t - kT) + \sqrt{E_2} \sum_k y_k(\mathbf{a}) h(t - kT) + n(t). \quad (1)$$

In (1) above, b_k is the kth member of the binary data stream $\mathbf{b}$ of the uncoded spot beam message, $\mathbf{a}$ represents the binary data stream for the global beam, $y_k(\mathbf{a})$ is the kth channel symbol of the global beam and is dependent upon $\mathbf{a}$ through the structure of the encoder, $h(t)$ is the impulse response of the channel, $n(t)$ is a Gaussian noise process of spectral power density $N_0/2$, and E_1 and E_2 are, respectively, the received pulse energy of the spot and global beam transmissions. We note that the b_k's are independent and equally likely to be ± 1, and that the y_k's can assume the values ± 1 but are not independent. We assume that intersymbol interference is absent.

A set of sufficient statistics[4] for detecting the $\mathbf{a}$ and $\mathbf{b}$ sequences is formed by the synchronous samples of $R(t)$ taken at the opening of the binary eye. One such sample is:

$$r_k = \sqrt{E_1}\, b_k + \sqrt{E_2}\, y_k + n_k. \tag{2}$$

We assume the various n_k's to be independent. From the samples (2), we form the log-likelihood function or path metric[4]

$$\Lambda(\mathbf{a},\mathbf{b}) = 2\sum r_k[\sqrt{E_1}b_k + \sqrt{E_2}y_k(\mathbf{a})] - \sum[\sqrt{E_1}b_k + \sqrt{E_2}y_k(\mathbf{a})]^2 \tag{3}$$

and decide upon those sequences $\hat{\mathbf{a}},\hat{\mathbf{b}}$ for which (3) is maximized.

The maximum-likelihood algorithm to perform optimum detection is similar to the Viterbi algorithm[5] and is illustrated by the state transition diagram of Fig. 4, drawn for a $K = 3$ convolutional code. The state is defined by the contents of the first two stages of the shift register, and knowledge of the starting state and the next bit entering the encoder uniquely determines the next state and the encoded channel symbols generated. We note that, unlike the ordinary Viterbi algorithm for rate $r = \frac{1}{2}$ codes, each transition between states can occur along four paths, rather than one, because two independent uncoded symbols are also

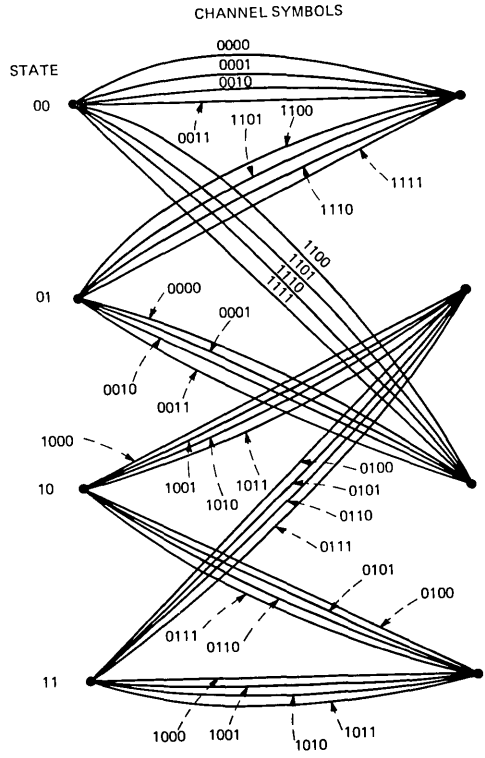


Fig. 4—State diagram for maximum-likelihood detection of interfering coded and uncoded signals. A $K = 3$, $r = \frac{1}{2}$ convolutional code is assumed. For each transition, the first two channel digits correspond to the coded symbols, and the second two correspond to the uncoded symbols.

generated during each epoch. The first two bits appearing along each branch correspond to the encoded channel symbols for that transition, and the second two digits correspond to one of the four possible two-bit sequences for the uncoded transmissions.

To perform maximum-likelihood detection, we note that, at each state, eight possible branches merge, and the partial path metric of one such merging branch must be the largest. The remaining seven paths then cannot be most likely because any succeeding additions to any one of these seven paths are valid additions to that one path exhibiting the greatest partial metric; succeeding additions, then, cannot cause the overall metric of any of these seven paths to exceed that of the path exhibiting greatest partial metric, and the seven paths having the smaller path metrics can be deleted from further consideration.

Thus, at each point in time, the four most likely paths (one leading to each state) and their associated partial metrics are known. During the next clock cycle, we determine the most likely of eight paths leading into each state by performing, for each of two initial states and for each of four branches for each initial state, the operation

$$\Lambda_n = \Lambda_{n-1} + \sum_{k=0}^{1} [2r_{2n-k} - \sqrt{E_1}b_{2n-k} - \sqrt{E_2}\,y_{2n-k}][\sqrt{E_1}b_{2n-k} + \sqrt{E_2}y_{2n-k}] \quad (4)$$

and saving the path and path metric of the largest for subsequent operations. The values of b_{2n-k} and y_{2n-k}, $k = 0,1$ to be substituted into (4) are determined from the state transition diagram, Fig. 4.

To perform true maximum likelihood detection, the most likely path leading into each state must be stored over the entire past. However, it has been shown that after 4 to 5 constraint lengths have elapsed, the oldest bits in all path memories are the same with a very high probability. Thus, we need to save only the most recent 4K through 5K of data for each state and, once in each epoch, the oldest bits in any one of four path memories can be outputted as detected data. We note that, unlike the ordinary Viterbi algorithm for which each path memory consists of a single rail of data, here we need to store three rails of data for each state. One rail contains the most likely source sequence **a** for the area coverage beam, and the second two contain the first and second source bits emitted each epoch for the uncoded spot beam sequence **b**.

The detector will commit an error for the first time at node n if the partial metric of some path which previously diverged from the correct path and remerges at node n is greater than that of the correct path. Some possible error events are shown in Fig. 5. We now calculate the probability of such an event.

Let **A** correspond to the spot and area coverage information sequence along the correct span, and let $\hat{\mathbf{A}}$ be those along the incorrect span. Then, the path metric difference is given by:

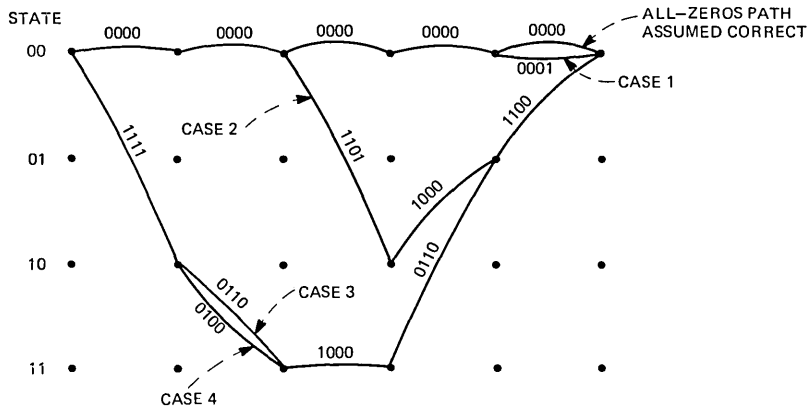


Fig. 5—Select error events for the maximum-likelihood detector.

$$\Lambda(\mathbf{A}) - \Lambda(\hat{\mathbf{A}}) = 2\sum r_k[\sqrt{E_1}(b_k - \hat{b}_k) + \sqrt{E_2}(y_k - \hat{y}_k)] - \sum[(\sqrt{E_1}b_k + \sqrt{E_2}y_k)^2 - (\sqrt{E_1}\hat{b}_k + \sqrt{E_2}\hat{y}_k)^2], \quad (5)$$

where the summation is performed over the unmerged span. Substituting (2) into (5) and recognizing that $b_k^2 = \hat{b}_k^2 = y_k^2 = \hat{y}_k^2 = 1$, we obtain

$$\Lambda(\mathbf{A}) - \Lambda(\hat{\mathbf{A}}) = A + n_{eq}, \quad (6)$$

where

$$A = 4\sum[\sqrt{E_1}\tilde{b}_k + \sqrt{E_2}\tilde{y}_k]^2 \quad (7)$$

and

$$n_{eq} = 4\sum n_k[\sqrt{E_1}\tilde{b}_k + \sqrt{E_2}\tilde{y}_k]. \quad (8)$$

In (7) and (8), we have used the nomenclature

$$\tilde{b}_k = \begin{cases} b_k & \text{if } \hat{b}_k = b_k \\ 0, & \text{otherwise.} \end{cases} \quad (9)$$

$$\tilde{y}_k = \begin{cases} y_k & \text{if } \hat{y}_k = y_k \\ 0, & \text{otherwise.} \end{cases} \quad (10)$$

The first event error probability P is equal to the probability that $\Lambda(\mathbf{A}) - \Lambda(\hat{\mathbf{A}}) < 0$. From (6) through (8), we conclude that

$$P = Q\left\{\sqrt{\frac{2\sum(\sqrt{E_1}\tilde{b}_k + \sqrt{E_2}\tilde{y}_k)^2}{N_0}}\right\}, \quad (11)$$

where Q is the complimentary error function. From this result, we now derive upper bounds on the bit error rate performance of the coded and uncoded transmission. We do the uncoded first.

Let the unmerged span be L channel digits long. We see from (11) that the first event error probability is dependent upon the correct sequence

along the unmerged span. For each possible error path of length L, we will determine the number of uncoded bit errors experienced along that path, and average (11) over all possible correct sequences. An upper bound on the average bit error rate for the uncoded transmission is then given by the summation over all possible incorrect paths, of the product of the number of bit errors experienced along a particular path and the average probability that the particular path has a metric exceeding that of the correct path.

Let the coded channel bits be different along the correct and incorrect paths in D symbols. Let the number of channel symbols for which an error occurs for both the coded and uncoded bits be denoted by r, and let the number of channel symbols for which an error occurs for the uncoded, but not for the coded, be denoted by s. Since the uncoded transmissions are equally likely to be ± 1, then along any L,D,r,s path, the coded and uncoded symbols may add or subtract, depending on the particular correct path. In $1/2^r$ of the paths, the correct symbols of the coded and uncoded transmissions will algebraically subtract over all r symbols. Similarly, in $\binom{r}{j}/2^r$ of the paths, there will be a subtraction in $(r-j)$ symbols and an addition in j symbols. In s symbols, $\tilde{b}_k^2 = 1$ and $y_k = 0$, while in $(D-r)$ symbols, $\tilde{b}_k = 0$ and $\tilde{y}_k^2 = 1$. There are $r+s$ errors committed in the uncoded transmission. Thus, averaging over all possible correct paths of the same L,D,r,s, we obtain the result that the average probability of error for each path of the same L and D for the uncoded transmission is given by

$$P_b = \frac{1}{2} \sum_{s=0}^{L-D} \sum_{r=0}^{D} \binom{L-D}{s} \binom{D}{r} \frac{r+s}{2^r} \sum_{j=0}^{r} \binom{r}{j} Q(r,s,j,D), \tag{12}$$

where

$$Q(r,s,j,D) \triangleq Q \left\{ \sqrt{\frac{2}{N_0}} [(r-j)(\sqrt{E_1} - \sqrt{E_2})^2 + j(\sqrt{E_1} + \sqrt{E_2})^2 + (D-r)E_2 + sE_1 \right\}. \tag{13}$$

The factor of $1/2$ appearing in front of (12) arises from the fact that two uncoded bits are transmitted per epoch.

Using the inequality that for $x \geq 0$, $y \geq 0$,

$$Q\{\sqrt{x+y}\} \leq Q\{\sqrt{x}\} e^{-y/2}, \tag{14}$$

we can overbound and simplify (12) and (13) to the following:

$$P_b \leq Q \left\{ \sqrt{\frac{2DE_2}{N_0}} \right\} e^{-E_1/N_0} [DXe^{E_1/N_0} + L(1+X) - D] \times [(1+X)^{D-1}(1+e^{-E_1/N_0})^{L-D-1}], \tag{15}$$

where

$$X \triangleq \frac{1 + e^{-4\sqrt{E_1E_2}/N_o}}{2} e^{-(E_1-2\sqrt{E_1E_2})/N_o}. \tag{16}$$

Finally, for a particular convolutional code, we use the code generating function matrix method of Viterbi[5] to identify all L,D paths for that code, and sum the contribution (15) for each such path over all possible paths. To this result must be added the contribution of the trivial case for which no coded errors occur (see Fig. 5, case 1). The contribution of these paths is simply

$$P_b = Q\left\{\sqrt{\frac{2E_1}{N_o}}\right\} + Q\left\{\sqrt{\frac{4E_1}{N_o}}\right\}. \tag{17}$$

The results of this exercise have been applied to the optimum $K = 7$, $r = \frac{1}{2}$ code,[6] and appear in Fig. 6. Plotted there is the uncoded bit error rate bound vs the required energy per information bit-to-noise ratio, e_b/N_o, for various ratios of interference to signal (E_2/E_1). Also plotted is the ideal, interference-free performance. We see here that as E_2/E_1 decreases below 2.5 dB, performance starts to improve. The utility of the maximum likelihood sequence estimation (MLSE) to detect uncoded transmission in the presence of coded area coverage interference is il-

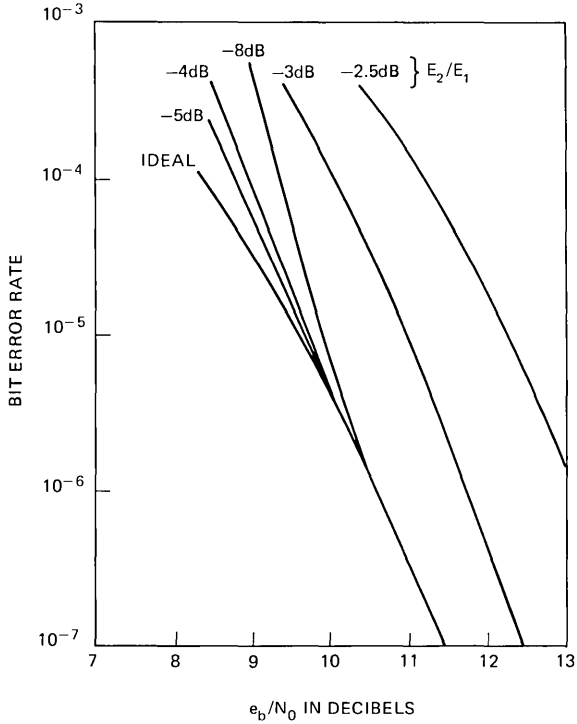


Fig. 6—Bound on uncoded bit error rate performance of maximum-likelihood detector vs e_b/N_o for select values of interference.

lustrated by the following example. Suppose $E_2/E_1 = -3$ dB. Then, if simple bit-by-bit detection is performed, an asymptotic degradation of about 10 dB from ideal would be expected. However, through use of the MLSE, the asymptotic degradation is about 1 dB.

We also see from these curves that, as E_2/E_1 decreases below about -8 dB, there is an apparent degradation in performance. This virtual result is caused by the bounding technique used, and is not experienced in practice. To see how this arises, we note that, as E_2/N_0 becomes small, all paths through the decoding trellis exhibiting a fixed number N of uncoded bit errors become equally likely. The contribution of each such path to the bit error rate bound is, however, summed, indicating a much higher bit error rate than would actually be encountered since only one such incorrect path could actually be selected at any node. For sufficiently small E_2/N_0, in fact, the bound no longer converges. To evaluate performance of the MLSE in the regime where the bound converges poorly, extensive simulation studies were performed and are shown in Fig. 7. These studies show that there is in fact a degradation in performance as E_2/E_1 decreases below -4 dB, but that the worst-case degradation of about 1 dB from ideal occurs for $E_2/E_1 = -10$ dB. As the in-

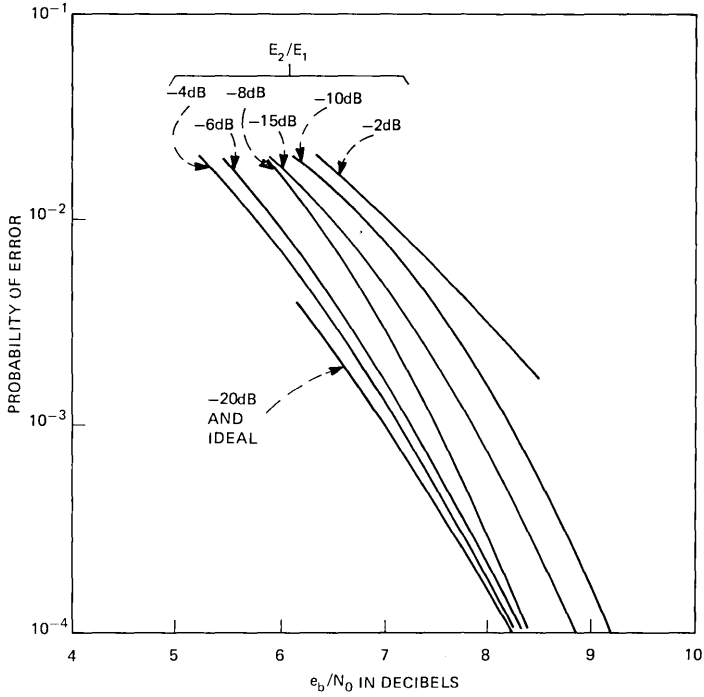


Fig. 7—Uncoded bit error rate performance of maximum-likelihood detector vs e_b/N_o obtained via simulation for select values of interference.

terference becomes smaller, performance begins to approach the ideal, interference-free case as intuitively expected.

We now study the performance of the encoded area beam. Again, let the unmerged span be L channel digits long and let the coded digits be different along the correct and incorrect paths in D channel symbols and N information symbols. Then, averaging overall possible combinations of the uncoded symbols, the average number of area beam bit errors incurred along any L,D,N path is given by:

$$P_b = N \sum_{s=0}^{L-D} \sum_{r=0}^{D} \binom{L-D}{s} \binom{D}{r} \frac{1}{2^r} \sum_{j=0}^{r} \binom{r}{j} Q(r,s,j,D), \tag{18}$$

where $Q(r,s,j,D)$ is given by (13). Invoking inequality (14), we obtain the bound:

$$P_b \leq N e^{-DE_2/N_0} (1 + e^{-e_1/N_0})^{L-D} (1 + X)^D, \tag{19}$$

where X is given by (16). Once again, we use the generating function matrix approach to determine the contribution of each incorrect path.

Results for the optimum $K = 7$ code appear in Fig. 8. Shown there is the bit error rate performance of the encoded area beam message vs e_b/N_o for select values of E_1/E_2, the interference-to-signal ratio. We see that, when E_1 becomes much greater than E_2, performance approaches the ideal, interference-free case since, under these conditions, the MLSE algorithm exploits the large difference between the signal and interference strengths to correctly decode the small signal. For $E_2 > E_1$, the bounding technique again suffers from poor convergence properties, and the results are meaningless. Again, extensive simulation studies were performed and are shown in Fig. 9. We see that, as expected, the ideal interference-free case is approached as E_1/E_2 becomes small.

For all values of E_1/E_2, the MLSE algorithm provides the best attainable performance. However, when E_1/E_2 becomes sufficiently small, the improvement possible via MLSE becomes negligible as shown by the plots of Fig. 10. These data were obtained experimentally and show the bit error rate performance of the ordinary Viterbi algorithm in the presence of a single bit-synchronous cochannel interferer. The ordinary Viterbi algorithm operates as though no interference was present and, unlike the MLSE algorithm, would be useless for $E_1 > E_2$. However, for $E_1 \ll E_2$, performance of the two are about the same, and the slight improvement possible via MLSE is not warranted in view of the additional complexity incurred.

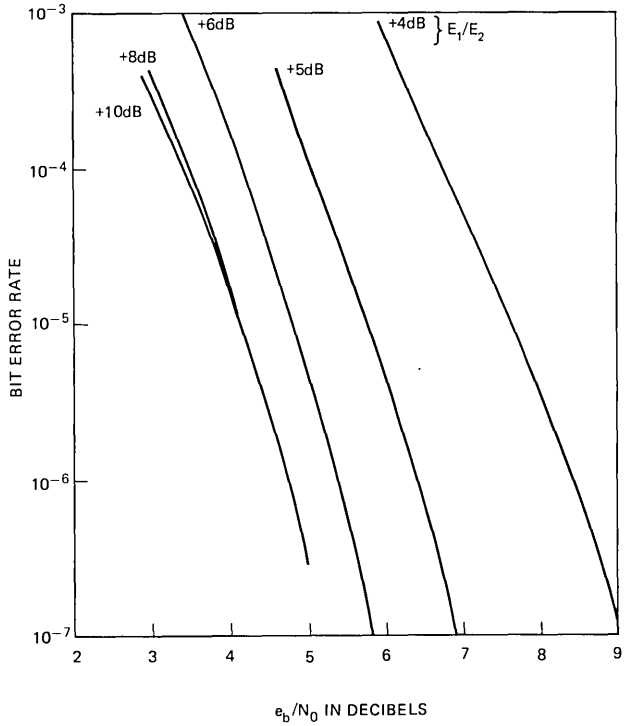


Fig. 8—Bound on coded bit error rate performance of maximum-likelihood detector vs e_b/N_o for select values of interference.

IV. APPLICATION

We now apply the results of the preceding section to the problem of reducing mutual interference between spot and area coverage beams sharing a common spectral band. Let the spot-beam radiation pattern be Gaussian-shaped and usable to its −3 dB contour. In the absence of cochannel interference, the e.i.r.p. of the coded global beam would be 8 dB lower than that of the spot beam at its −3 dB contour for the same system outage and bit error rate (BER) performance. This 8-dB factor can be broken down into a 3-dB component, since the information rate of the global beam is half that of the spot beam, plus a 5-dB component representing the coding gain of a $K = 7$, $r = 1/2$ convolutional code. Suppose we set E_2/E_1 at the 3-dB contour of the spot beam at −8 dB. Then, throughout the spot-beam coverage area, $-11 \text{ dB} \leq E_2/E_1 \leq -8 \text{ dB}$. From Fig. 6, we see that, over this range, the BER performance of the uncoded spot-beam message is degraded by at most 1 dB if MLSE is employed. By contrast, if bit-by-bit detection of the spot-beam message were employed, the degradation would be between 2.9 dB and 4.4 dB.

Let the e.i.r.p. of both the spot and area coverage beams rise by 1 dB.

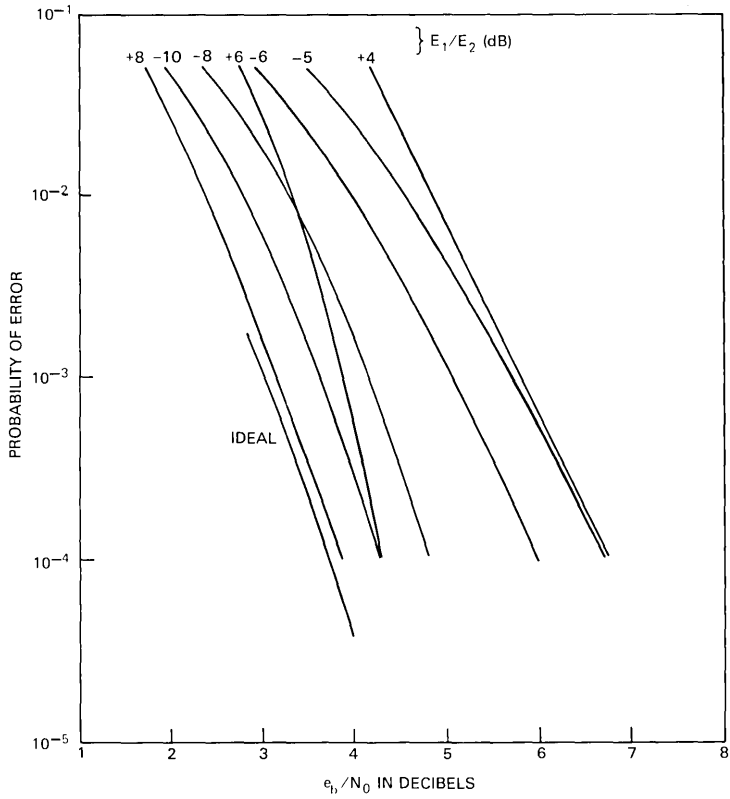


Fig. 9—Coded bit error rate performance of maximum likelihood detector vs e_b/N_o obtained via simulation for select values of interference.

Then, throughout the spot-beam coverage region, the BER performance is at least as good as that obtained in the absence of interference with 1 dB less power. From Fig. 8, we see that, beyond the −3 dB contour of the spot beam, we can communicate via the area coverage beam in conjunction with MLSE with at most 1-dB degradation from the ideal interference-free situation provided $E_1/E_2 > 5.5$ dB. Finally, from Fig. 10, we see that we can use the area beam with the ordinary Viterbi algorithm provided $E_1/E_2 < -12$ dB. From these observations, we can construct the plot of Fig. 11, which shows the one-dimensional radiation patterns of a spot beam and the area beam and the usable regions for the spot and area coverage beams in the vicinity surrounding a spot beam. Implicit in this illustration is the fact that the e.i.r.p. of both the spot and area beams is increased by 1 dB to provide the same grade of service as possible with 1 dB less power in the absence of cochannel interference. We see that communication via the spot beam, in conjunction with MLSE, is employed out to the −3 dB contour of the spot beam. From $\theta = \theta_{3\text{dB}}$ out to $\theta = 1.4\theta_{3\text{dB}}$, we can communicate via the area beam, even though

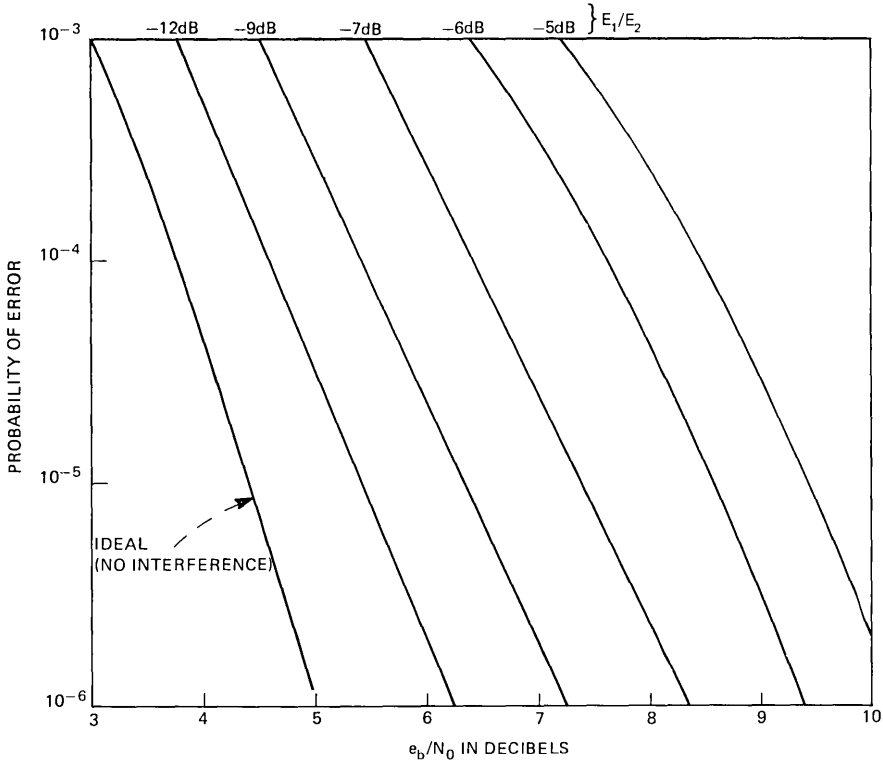


Fig. 10—Measured bit error rate performance of a $K = 7, r = 1/2$ convolutional code vs e_b/N_o for select values of interference. The interference is bit-synchronous with the encoded channel bits, and the ordinary Viterbi algorithm with soft (3-bit) quantization is employed.

the interference is stronger than the desired signal. Between $\theta = 1.4\theta_{3dB}$ and $\theta = 2.75\theta_{3dB}$, the performance degradation of the area beam exceeds the allotted 1 dB, and the desired grade of service cannot be provided. This region, then, is blacked out. Finally, for $\theta > 2.75\theta_{3dB}$, communication via the global beam is again possible.

Thus, through utilization of an area coverage beam in conjunction with channel coding and MLSE, the blackout region of a multiple spot-beam communication satellite is reduced from the entire region not serviced by any spot beam to a thin annular ring surrounding each spot beam. There is no sacrifice in the capacity of the spot beams, and the power penalty is 1 dB for all beams.

Let us now consider a specific example. We assume the existence of 10 spot beams, half of which employ one polarization and half the orthogonal polarization. In the absence of interference, each spot beam transponder uses a 3-watt final power amplifier, and the difference between the spot and area beam antenna gains is 20 dB. Suppose we deploy

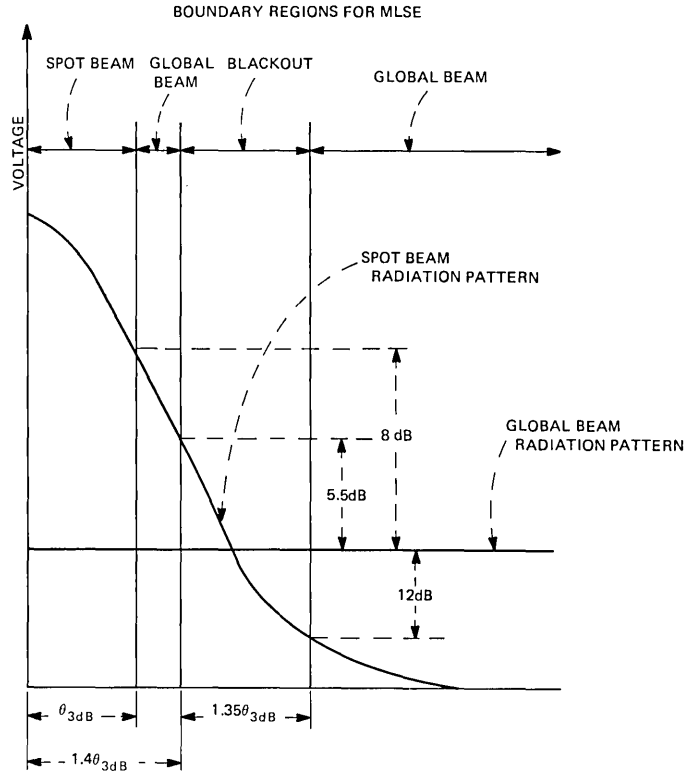


Fig. 11—One-dimensional plot showing the usable regions attainable via MSLE for a hybrid spot-area beam satellite system employing a $K = 7, r = 1/2$ code for the area beam transmission. The spot beam antenna pattern is Gaussian shaped.

a single-area beam transponder employing one of the two polarizations; the capacity of this beam is one-half that of a spot beam, and a $K = 7$, $r = \frac{1}{2}$ code is employed. In the absence of interference, the RF power required of the area beam would be $20 - 8 = 12$ dB higher than any spot beam. The total required RF power for the hybrid system outlined above is then

$$P = 1.25 \times [3 \times 10 + 47.5] = 97 \text{ watts.} \quad (20)$$

By contrast, if we employ the band-splitting technique described in Section II, we would need 6 dB more power for each spot beam, and the power required for the area beam would be 17 dB higher than that required for the spot beam in the absence of interference, since coding is not employed. The total power, then, would be

$$P = 4 \times 3 \times 10 + 50.1 \times 3 = 270 \text{ watts.} \quad (21)$$

Considering a 30-percent efficiency for the final TWT, the total dc power

required via coding is 323 watts, while that needed for the alternative band-splitting approach is 900 watts.

Since, through use of coding, the dc power required for an area beam is only 158 watts, we might consider deploying a second area beam using the orthogonal polarization. Then, not only do we double the capacity into the outlying region, but we also eliminate the blackout region, since each spot beam is used in only one polarization. Area coverage communication to the blackout region of one polarization can thereby be provided in the second polarization. The dc power required for this approach is 442 watts.

V. MAXIMUM-LIKELIHOOD ALGORITHM WITH FIXED PHASE SHIFT

In Section III, we derived a maximum-likelihood algorithm which allows joint area and spot-beam coverage sharing a common spectral band whenever there is no carrier phase shift difference between the area and spot beam transmissions. We now derive the proper algorithm for use when there is a fixed phase shift difference, θ. During the kth clock cycle, the spot beam source emits two bits, $b_{1,k}$ and $b_{2,k}$, and the area beam source emits a single bit a_k and two encoded channel bits $y_{1,k}(\mathbf{a})$ and $y_{2,k}(\mathbf{a})$. The data $b_{1,k}$ and $b_{2,k}$ are modulated onto a carrier via 4ϕ-PSK, as are $y_{1,k}$ and $y_{2,k}$. Thus, we transmit:

$$R(t) = \sqrt{E_1}b_{1,k}\cos(\omega t + \theta) + \sqrt{E_1}b_{2,k}\sin(\omega t + \theta) + \sqrt{E_2}y_{1,k}\cos\omega t + \sqrt{E_2}y_{2,k}\sin\omega t. \quad (22)$$

The receiver locks onto the phase of the encoded area beam and, during the kth clock cycle, the receiver observes, after coherent demodulation, the two test statistics:

$$r_{1,k} = \sqrt{E_2}y_{1,k} + \sqrt{E_1}b_{1,k}\cos\theta + \sqrt{E_1}b_{2,k}\sin\theta + n_{1,k} \quad (23)$$

$$r_{2,k} = \sqrt{E_2}y_{2,k} - \sqrt{E_1}b_{1,k}\sin\theta + \sqrt{E_1}b_{2,k}\cos\theta + n_{2,k}. \quad (24)$$

The path metric now takes the form:

$$\Lambda(\mathbf{a},\mathbf{b}) = \sum_k [r_{1,k}(\sqrt{E_2}y_{1,k} + \sqrt{E_1}b_{1,k}\cos\theta + \sqrt{E_1}b_{2,k}\sin\theta) + r_{2,k}(\sqrt{E_2}y_{2,k} - \sqrt{E_1}b_{1,k}\sin\theta + \sqrt{E_1}b_{2,k}\cos\theta)] - \sqrt{E_1E_2}[y_{1,k}(b_{1,k}\cos\theta + b_{2,k}\sin\theta) - y_{2,k}(b_{1,k}\sin\theta - b_{2,k}\cos\theta)]. \quad (25)$$

As before, we define the state of the encoder by the contents of the first $K-1$ stages of its shift register, and each state can be accessed via eight paths. Along each path, we compute the partial metric:

$$\Lambda_k(\mathbf{a},\mathbf{b}) = \Lambda_{k-1}(\mathbf{a},\mathbf{b}) + \sqrt{E_2}r_{1,k}y_{1,k} + \sqrt{E_2}r_{2,k}y_{2,k} + \sqrt{E_1}(r_{1,k} - \sqrt{E_2}\,y_{1,k})(b_{1,k}\cos\theta + b_{2,k}\sin\theta) - \sqrt{E_1}(r_{2,k} - \sqrt{E_2}\,y_{2,k})(b_{1,k}\sin\theta - b_{2,k}\cos\theta), \quad (26)$$

and save the path and metric of the larger. Thus, with a fixed known phase shift, maximum-likelihood decoding is also possible.

VI. CONCLUSIONS

In multiple spot-beam communication satellite systems, it is often desirable to provide service to remote areas not covered by any spot beam. This additional service should neither diminish the capacity of the spot beams nor cause a severe downlink power penalty. We considered deployment of an area beam transponder, in addition to the fixed spot beams, and saw that satisfaction of the above requirements implies considerable downlink cochannel interference at all ground stations located in the vicinity of any spot beam. The use of binary convolutional codes for the area beam transmission was shown to greatly curtail the performance degradation resulting from this cochannel interference and also reduce the prime power requirements of the area beam transponder.

A maximum-likelihood algorithm was derived to optimally detect either the uncoded spot beam transmission or the coded area beam transmission, and performance of this algorithm was evaluated. Use of this algorithm was shown to provide for reliable spot-beam communication in the presence of cochannel interference. It is also possible to reliably communicate via the global beam in the presence of a much stronger spot-beam interference. These results were then applied to a scenario in which interference was reduced on the uplink via the simple technique of band-splitting between the area and spot beams. Such a technique is unsuitable for the downlink because of the power penalty incurred. On board, the uplink bits are regenerated and switched into the appropriate downlink beam, and a $K = 7$, $r = \frac{1}{2}$ code is employed for the downlink area beam. Results show that the degradation from cochannel interference is contained to be less than 1 dB over the entire service area except for a thin annular ring surrounding each spot beam. Traffic originating within or destined for these blackout rings might be backhauled to the nearest serviceable region, or else a second area beam, employing the dual polarization, might be deployed such that, for any given spot beam, the blackout region is contained to only one polarization. Since the spot beams use both polarizations to minimize interference among themselves, the MLSE algorithm must still be used at all spot-beam ground stations to provide spot-beam service with minimal performance degradation.

The satellite prime power demands to satisfy RF radiated power requirements were evaluated and shown to be within the capability of the Thor-Delta class. Thus, the use of spot and area coverage beams, sharing a common spectral band, in conjunction with channel coding techniques, appears to be an acceptable method for providing universal service via high-capacity digital switching satellites of the future.

VII. ACKNOWLEDGMENTS

The author wishes to thank his colleagues, D. O. Reudink and Y. S. Yeh, for their stimulating discussions and contributions, and also Mrs. Phyllis Arnold who wrote the programs for the simulation studies.

REFERENCES

1. L. C. Tillotson, "A Model of a Domestic Satellite Communication System." B.S.T.J., *47,* No. 10 (December 1968), pp. 2111–2137.
2. R. Cooperman and W. G. Schmidt, "Satellite Switched SDMA and TDMA Systems for Wideband Multi-Beam Satellite," ICC Conference Record, 1973.
3. A. S. Acampora, D. O. Reudink, and Y. S. Yeh, "Spectral Re-Use in 12 GHz Satellite Communication Systems," ICC Conference Record, 1977.
4. H. L. Van Trees, *Detection, Estimation, and Modulation Theory,* Part I, New York; Wiley, 1968.
5. A. J. Viterbi, "Convolutional Codes and Their Performance in Communication Systems," IEEE Trans. Comm. Tech., *COM-19* (October 1971), pp. 751–772.
6. K. S. Gilhousen, "Coding Systems Study for High Data Rate Telemetry Links," Linkabit Corporation, San Diego, California, January 1971.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 9, September 1978
Printed in U.S.A.

Reliability of a Microprocessor-Based Protection Switching System

By G. S. FANG

(Manuscript received July 7, 1977)

High-capacity transmission systems usually include one or more hot spares for protection. When a regular transmission channel fails, its signal is rapidly transferred to the spare channel under the control of protection switching circuits so that there is little signal degradation or interruption. This paper studies the reliability of a microprocessor-based terminal protection switching system. Some new and interesting behavior patterns for transmission systems with automatic protection switching are revealed. Also, some new memory self-checking algorithms are presented which increase the capability of microprocessor system fault recognition.

I. INTRODUCTION

In high-capacity transmission systems, any failure may affect a large number of message circuits. Such systems usually include one or more hot spares to increase system reliability. When a regular transmission channel fails, its signal is rapidly transferred to the spare channel under the control of protection switching circuits so that there is little signal degradation or interruption. This paper studies the reliability of a microprocessor-based terminal protection switching system (TPSS). The specific transmission facility under consideration is the L5E coaxial cable analog system, which is an expanded version of the L5 system.[1] The L5E multiplex equipment, or multimastergroup translators (MMGT), carry up to eight mastergroups, or 4800 telephone circuits. The TPSS will automatically switch into service a protection MMGT in the event of a failure of any one of up to 20 MMGTs.

Reliability theory has been studied by numerous authors,[2,3] and almost every Bell System transmission facility with automatic protection switching has been the subject of at least one reliability study.[4,5] The present analysis was undertaken for several reasons. First, many simplifying assumptions were made in the previous studies. Not all the

effects of the reliability of the switch, the protection switching control circuit, and the monitor circuit failures were taken into account. Second, in most cases, exponentially distributed restoration time has been assumed. This means that the probability of restoration at any instant after a failure is assumed to be independent of how much time has already been spent on restoring the failure. This assumption is rarely true in high-capacity transmission systems. Third, only steady-state analyses were made. A system with hidden failures will not reach its steady state in its lifetime. Fourth, a microprocessor-based protection switching control circuit has not been studied in such detail before. Finally, past experiences have shown that maintenance-induced service outages contribute to a very big share of the total outage time. This study also tries to take these outages into consideration.

With the MMGT system as an example, the present study attempts to analyze the same reliability problem in more detail and with less restrictive assumptions. Section II describes the protection switching arrangement. Section III explains the specific approaches used in this paper. Section IV presents the results graphically to emphasize the various reliability trends. Section V summarizes the conclusions obtained. Appendix A investigates some new microprocessor self-checking algorithms and Appendix B presents the derivations.

II. MMGT PROTECTION SWITCHING SYSTEM DESCRIPTION

Figure 1 is a simplified MMGT-system block diagram which illustrates the $1 \times n$ protection switching arrangement. There is one protection channel in each direction of transmission. Under the command of the microprocessor, each protection channel protects up to n regular channels, where n is equal to 20 in the TPSS. The same processor is used to control the switching actions of both directions of transmission. The switches are all solid-state devices, and their normal states are indicated in the figure. The crucial output switches are dual-powered. Parts of the output switch are designated the through switch and the substitute switch for later reference.

When there is no alarm from the various regular pilot detectors, the processor exercises the input switches for each channel sequentially to detect possible protection failures. In the event of a failure of one of the regular channels, the corresponding pilot detector sends an alarm to the processor. If the protection channel is available, the processor will first switch the input signal through the input switches to feed the protection channel. Whether the protection detector indicates a good signal or not, the processor will complete the 1×2 output switch. The regular detector is now monitoring the signal supplied by the protection channel via the output substitute switch. If the regular detector still alarms after the protection switch, the switching action will be reversed. The 1×2 output

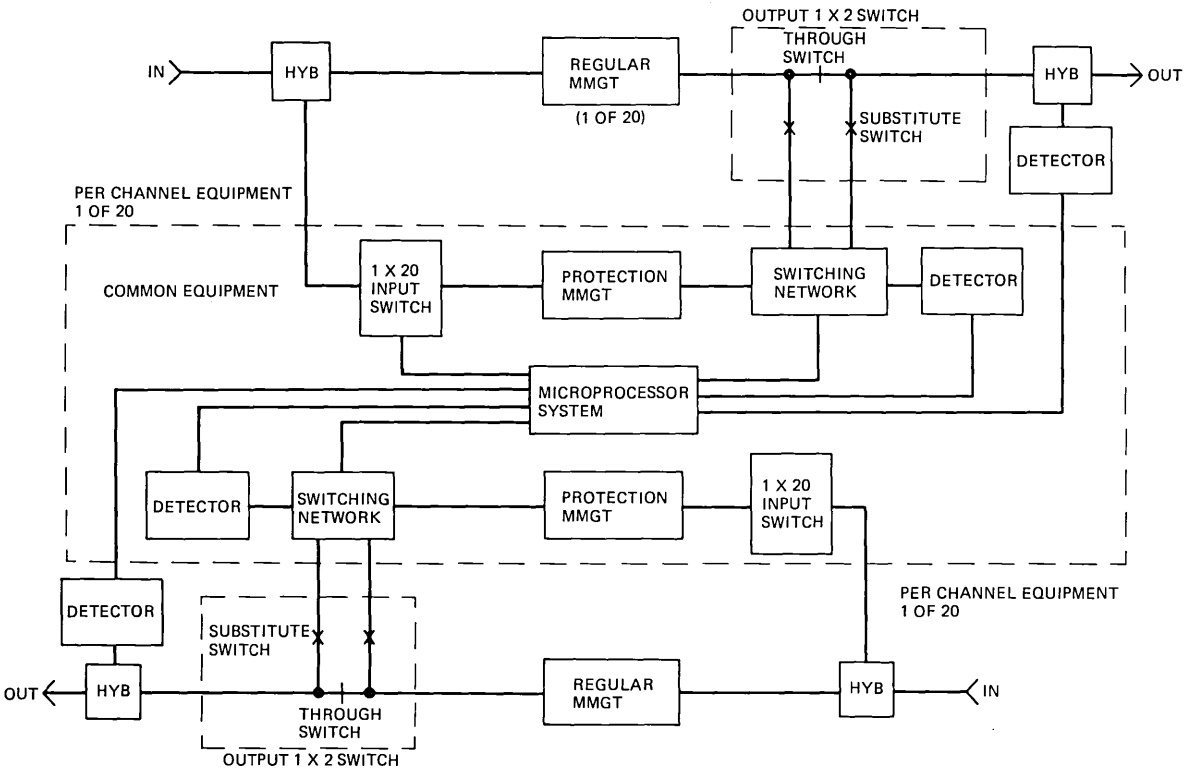


Fig. 1—TPSS block diagram.

switch will be deactivated and the input switch released. If the regular detector stops alarming after the output switching, a successful protection switch has been made, and the protection detector is monitoring the failed regular channel. When the failed channel is repaired, the protection detector will see a good signal, and the switches will return to their normal states. The protection channel is then free to service another regular channel failure.

Service outages can occur in many ways. In addition to multiple transmission failures, they can also be generated by the failures of the detectors, the switches, or the microprocessor system. The various failure modes are taken into account in later derivations.

III. APPROACHES

Two reliability measures of interest in transmission systems are used in this study. The first measure is the probability of service outage due to equipment failures. This probability translates directly to the system outage time per year and is the most commonly used figure of merit in determining transmission system reliability. The second measure is the probability of having maintenance activities going on. This measure will be abbreviated as the probability of activity. It is believed to be closely related to the probability of having maintenance-induced outages. This probability of activity is greater than the probability of having alarms because there are failures that cannot be detected locally. For instance, if the pilot detector for a failed regular channel is stuck to the state of no alarm, the failure can only be detected by downstream offices. Thus there may be maintenance activities in an office but no alarm. The probability of activity is less than the probability of having failures because there are undetectable failures such as the breakdown of an output substitute switch. A reliable system should have a small probability of outage and a small probability of activity.

Two additional criteria are used to measure the effectiveness of the overall protection plan. The improvement factor (IF) is defined as the ratio of the probability of outage without protection switching to that with protection switching. The activity factor (AF) is defined as the ratio of the probability of activity with protection switching to that without protection switching. These definitions agree with the common notion that an effective protection plan should provide more improvement and less activity. Thus, a better protection system has a bigger IF and a smaller AF. The activity factor is always greater than one.

The probabilities discussed above are derived under the assumptions that the various failures are statistically independent and the failure rates are constant. These are very simple assumptions considering the complexity of the problem. The assumption of statistical independence is made to avoid estimating conditional failures, although there is

probably dependency between the through switch and the substitute switch. The constant failure rate implies exponentially distributed failures, i.e., any working item is as good as new. This is a reasonable assumption for solid-state devices after the initial "burn-in" period. Notice that no distributional assumption is made on the restoration time. Based only on the failure rates and the restoration times of the components of the system, the various probabilities are derived from the basic definitions of conditional probability. Not only does this approach require little mathematical background, but the result is more general and more accurate than the usual method of Markoff chain or birth-and-death stochastic processes,[2,3] which assume that both failure and restoration times are exponentially distributed.

IV. DETAILED RESULTS

Table I introduces the notations and gives the estimated failure rates in FITS (number of failures per component per 10^9 hours), restoration times in hours, and the availabilities of the various components. The restoration time is the sum of the detection time and the equipment replacement time. The mean value of the replacement time t is assumed to be 1 hour. Some failure rates are expressed in terms of other failure rates to show their relative dependence. This is necessary in later parameter sensitivity studies. The failure of a substitute switch can only be detected when its use is called for. Thus, its detection time is the mean time between transmission failures of its corresponding channel, i.e., $1/(\lambda_r + \lambda_t + \lambda_0)$. The same is true for the detection time of a regular detector, except that the assumed probability that a failed detector gives a no-alarm indication is 1/4. In both cases, the equipment replacement time is ignored since it is small compared with the detection time.

The detection times of the hidden CPU (central processing unit) and EROM (erasable read-only memory) failures should also be similarly calculated. However, the failure of the regular channels to be exercised sequentially should provide local craftspeople with the indication that something is wrong. Therefore, the detection times are assumed to be 24 hours. The availability[3] of an item is the probability that the item is working. It is a function of time with an initial value of one and with a steady-state value equal to the mean time to failure divided by the sum of the mean time to failure and the mean restoration time. If a component has a short failure detection time, the transient portion in its availability value vanishes quickly, and the steady-state theoretical availability approximates the actual availability very well. For example, the steady-state availability of the regular channel is $p_r = 1/1.000001$. The reliability function of the regular channel is $e^{-10^{-6}t}$. It takes only 1 hour for the reliability function to reach its steady-state availability value.

Table I — Estimated failure rates

	FITS	Mean Restoration Time (hr)	Availability
Regular channel	$\lambda_r = 1000$	$\mu_r = t$	$p_r = \frac{1}{1 + \lambda_r \mu_r}$
Through switch	$\lambda_t = 150$	$\mu_t = t$	$p_t = \frac{1}{1 + \lambda_t \mu_t}$
Output switch	$\lambda_0 = \frac{1}{3}\lambda_t$	$\mu_0 = t$	$p_0 = \frac{1}{1 + \lambda_0 \mu_0}$
Substitute switch	$\lambda_s = \frac{2}{3}\lambda_t$	$\mu_s = \frac{1}{\lambda_r + \lambda_t + \lambda_0}$	$p_s = \frac{1}{1 + \lambda_s \mu_s} + \frac{\lambda_s}{(\lambda_s + \mu_s^{-1})^2 T}[1 - e^{-(\lambda_s + \mu_s^{-1})T}]$
Regular detector	$\lambda_d = 300$	$\mu_d = \frac{1}{4}\frac{1}{\lambda_r + \lambda_t + \lambda_0}$	$p_d = \frac{1}{1 + \lambda_d \mu_d} + \frac{\lambda_d}{(\lambda_d + \mu_d^{-1})^2 T}[1 - e^{-(\lambda_d + \mu_d^{-1})T}]$
Protection detector	$\lambda_D = 300$	$\mu_D = t$	$p_D = \frac{1}{1 + \lambda_D \mu_D}$
Protection channel	$\lambda_p = \lambda_r + 4\lambda_s + 100$	$\mu_p = t$	$p_p = \frac{1}{1 + \lambda_p \mu_p}$
CPU	$\lambda_c = 500$	$\mu_c = 24 + t$	$p_c = \frac{1}{1 + \lambda_c \mu_c}$
EROM	$\lambda_e = 300$	$\mu_e = \mu_c$	$p_e = \left(\frac{1}{1 + \lambda_e \mu_e}\right)^4$
RAM	$\lambda_a = 400$	$\mu_a = t$	$p_a = \left(\frac{1}{1 + \lambda_a \mu_a}\right)^2$

These arguments do not hold for failures requiring long detection times. For instance, the mean time to failure and the mean restoration time of a substitute switch are in the order of hundreds of years, while the life span of the equipment is expected to be only 40 years. To obtain an appropriate availability in such cases, one would observe that the restoration time of the substitute switch is exponentially distributed. This is due to the fact that the replacement time is ignored and the detection time depends on the transmission failures which are exponentially distributed. Thus the availability function can be derived explicitly as

$$A_s(t) = \frac{1}{1 + \lambda_s \mu_s} + \frac{\lambda_s}{\lambda_s + \mu_s^{-1}} e^{-(\lambda_s + \mu_s^{-1})t}.$$

The availability p_s given in Table I is the $A_s(t)$ averaged over the life span T of the equipment. The availability of the detector p_d is obtained similarly. The availability expressions of the EROM and the RAM reflect the use of 4 EROMs and 2 RAMs in the TPSS.

To gain insight and to study the sensitivity of the derived probabilities to the estimated failure rates and restoration times, the various estimated parameters are varied one at a time to show the system reliability trends. The results are presented graphically in the figures. In each figure, the solid line corresponds to the ordinate at the left and the dotted line to that at the right.

Figures 2 through 7 present the variations of the outage and the activity probabilities as functions of the regular channel, the detector, the switch, the CPU, the EROM, and the RAM failure rates, respectively. Most of the curves are almost linear because, for the small failure rates of interests, they are still in their linear regions. As far as the probability of outage is concerned, undetectable failures are the most damaging. The hidden detector and the substitute switch failures contribute to the bigger slopes in Figs. 3 and 4. Increasing the microprocessor system failures adds very little to the outage probability, as can be seen from Figs. 5 to 7. The probability that has the fastest increase is the switch failure rates because there are so many switches in the system. Figure 8 indicates that service outage can increase substantially if the replacement time for failed equipment is long. Figure 9 shows the effect of varying the detection time of the hidden microprocessor failure. Neither the outage nor the activity probability is sensitive to the detection time. Figure 10 shows the effect of varying the number of regular channels equipped. The discrete points in the figure are connected to show the almost linear trends. When the system is fully loaded, i.e., $n = 20$, there are about 2 minutes of service outage each year due to equipment failures and there is about half an hour of maintenance activities. It should be emphasized that the curves present the right trends

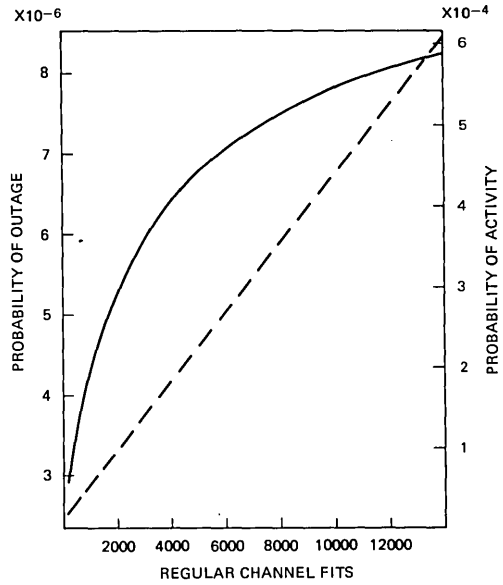


Fig. 2—Probabilities of outage and activity as functions of regular channel failure rate.

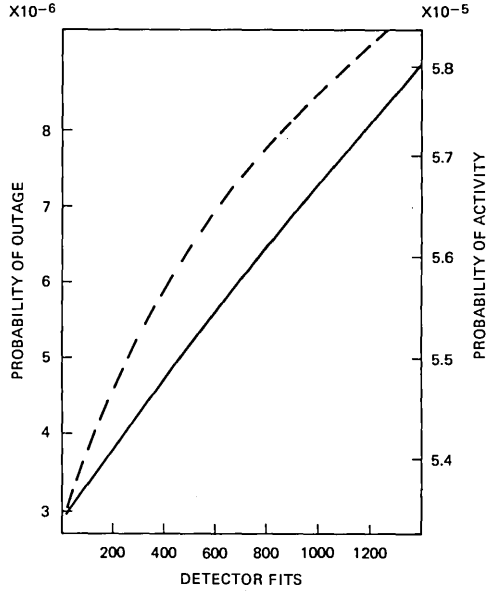


Fig. 3—Probabilities of outage and activity as functions of detector failure rate.

rather than numerical accuracy. From Fig. 2, if the failure rate of the regular channel is increased by ten times, there will be 4 minutes of outage and 4 hours of activity each year. Figure 10 shows the two

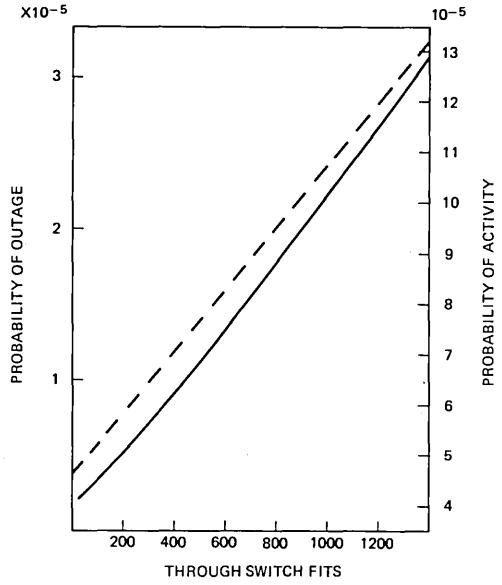


Fig. 4—Probabilities of outage and activity as functions of through switch failure rate.

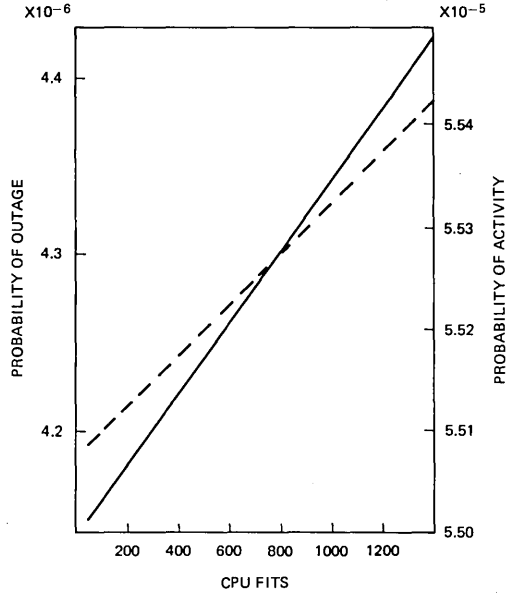


Fig. 5—Probabilities of outage and activity as functions of CPU failure rate.

probabilities as functions of the number of regular channels. The discrete points are connected to indicate trends. For terminal circuits which usually have small failure rates, there is scarcely any need for a second

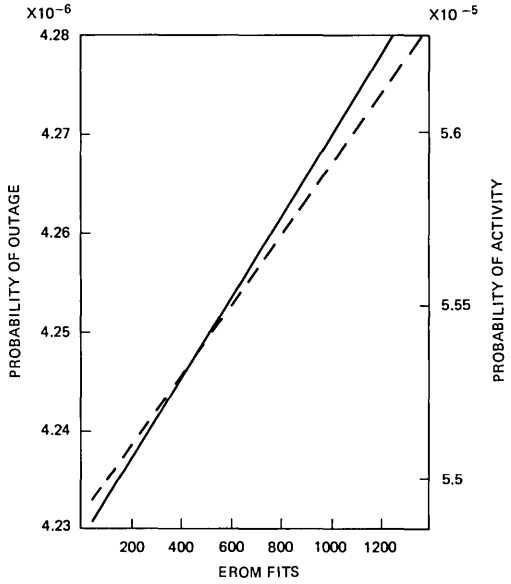


Fig. 6—Probabilities of outage and activity as functions of EROM failure rate.

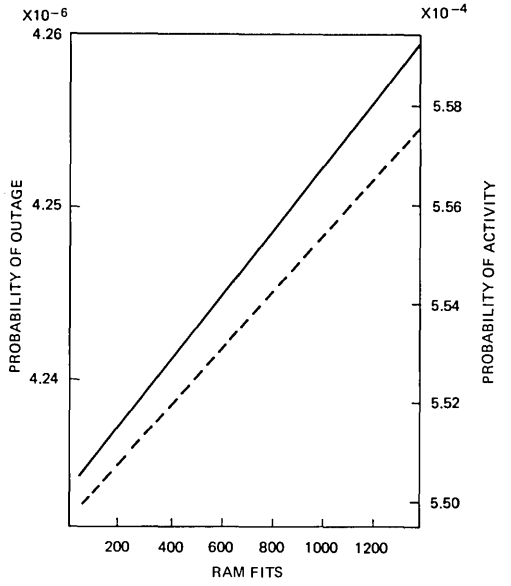


Fig. 7—Probabilities of outage and activity as functions of RAM failure rate.

protection channel even when the number of regular channels is large.

A system without protection switching has only the regular channels and their corresponding detectors to indicate alarms. The switches and

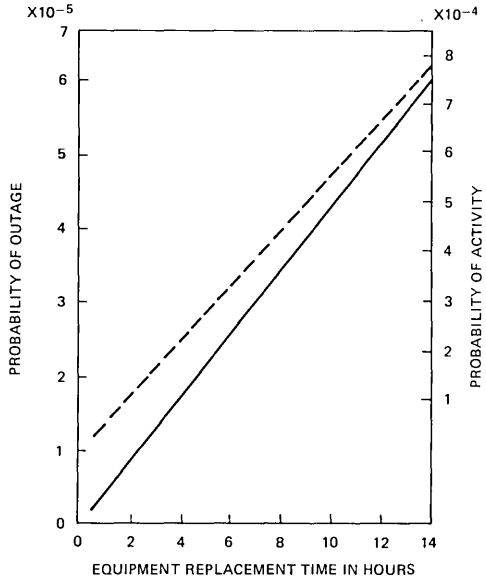


Fig. 8—Probabilities of outage and activity as functions of equipment replacement time.

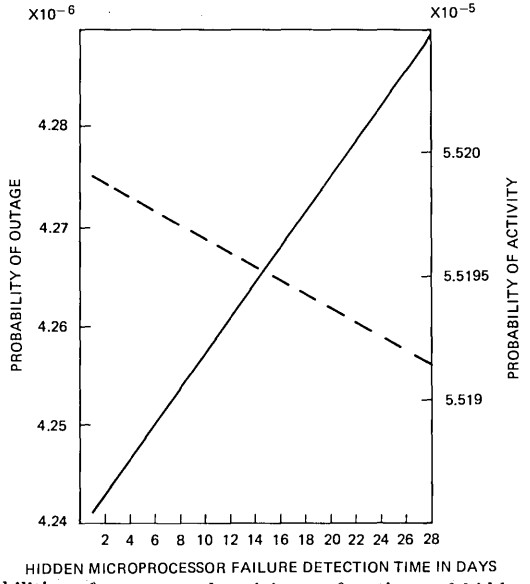


Fig. 9—Probabilities of outage and activity as functions of hidden microprocessor failure detection time.

the microprocessor devices are not required. Thus there is definitely less activity in the maintenance offices. Figure 11 shows the trend that, for small regular channel failure rates, the IF can be less than unity, i.e.,

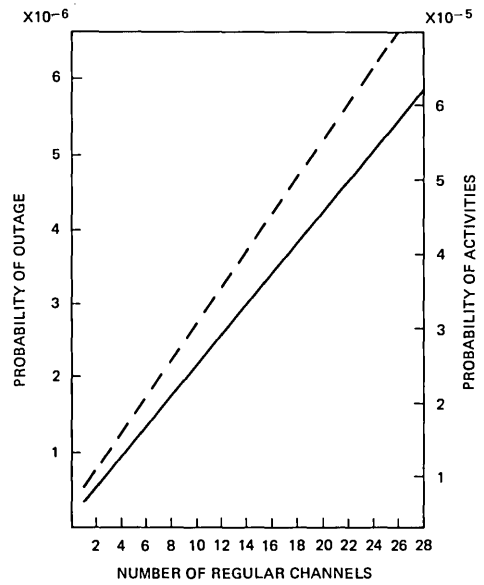


Fig. 10—Probabilities of outage and activity as functions of number of regular channels.

having protection switching actually causes more service outage. This is true when the failure rate of the regular channel is small compared with those of the protection switching circuits. Furthermore, protection switching generates many more activities at low regular channel failure rates. Figure 12 amplifies this fact by examining the 1 × 1 configuration. The IF is so small and the AF is so big that implementation of a 1 × 1 protection plan is questionable at low failure rates. Figure 13 gives the variations of the two factors with detector failure rates. Since detector failures have little effect on the outage probability of an unprotected system, the IF decreases with increasing detector failure. The interesting shape of the AF curve is due to the relatively rapid increase in the probability of activity for an unprotected system when the detector failure rates are small. This behavior is unique to the variation of the detector failure rate because an unprotected system is equipped only with the transmission channels and the detectors.

Figure 14 again indicates the important role played by the output switch. If its failure rate is high enough, the IF can reduce to less than unity. With a perfect switch, the outage of a protected system can be hundreds of times less than that of an unprotected system. The curves showing the two factors as functions of the CPU, the EROM, and the RAM failure rates are not given here. These curves can be simply deduced from Figs. 5 to 7 because the various probabilities of an unprotected system are independent of microprocessor failures. Similarly, the factors involving hidden microprocessor failure restoration time can be obtained

from Fig. 9. Figure 15 shows that both the IF and the AF are not very sensitive to how long it takes to replace failed equipment. Figure 16 varies the number of regular channels. It indicates that more than 10 regular

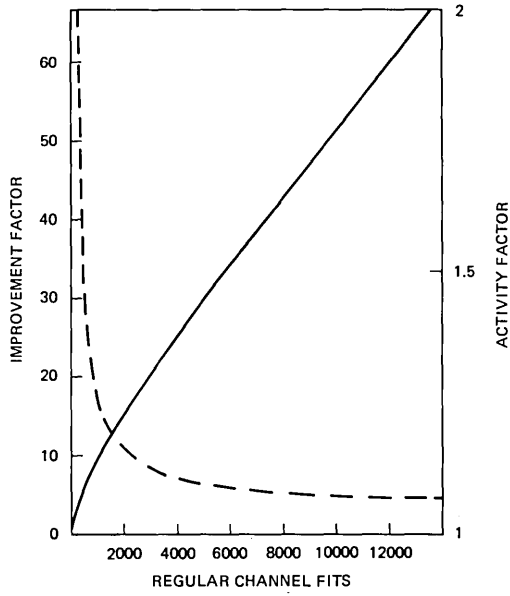


Fig. 11—Improvement and activity factors as functions of regular channel failure rates.

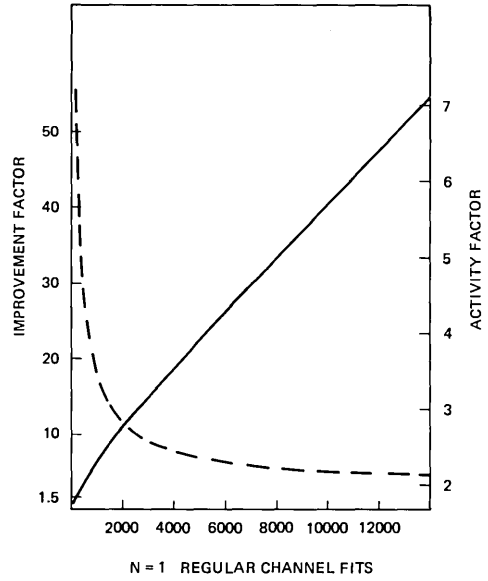


Fig. 12—Improvement and activity factors as functions of regular channel failure rates.

channels should be used to take advantage of the protection switching arrangement.

Figure 17 exhibits an interesting behavior of general protection

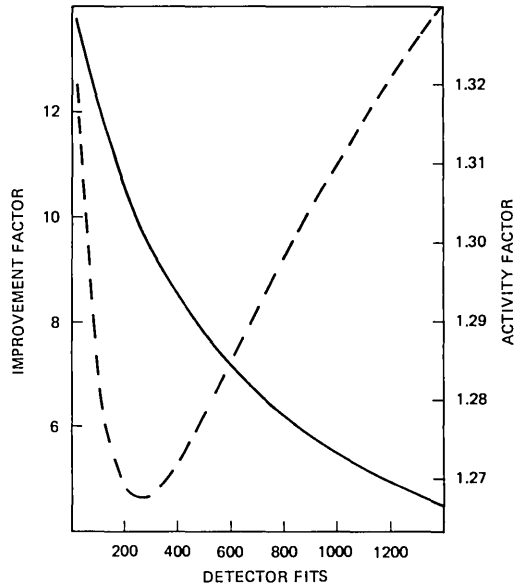


Fig. 13—Improvement and activity factors as functions of detector failure rates.

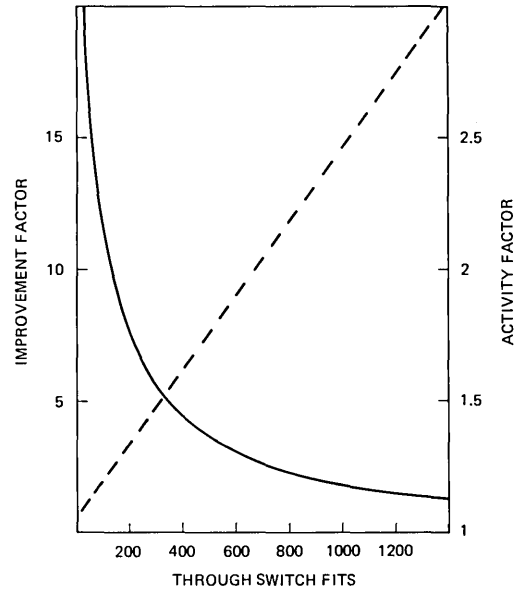


Fig. 14—Improvement and activity factors as functions of through switch failure rates.

switching systems. As the failure rate of the regular channel increases, the IF increases from less than one to a maximum and then starts to decrease. When the failure rate becomes very large, the outage probability is close to 1 with or without protection switching. Thus the IF approaches 1 eventually. The maximum IF shown in the figure occurs at around 150,000 FITS. Although it is unlikely for a terminal multiplexer to possess so high a failure rate, a line transmission system with many cascading repeaters may very well have a failure rate of this order. Therefore, whenever a line protection switching system is planned, the reliability should be studied to determine the length of the protection span so that the IF does not fall in its decreasing region. Of course, the outage probability should also be taken into account to meet any prescribed service objectives.

V. CONCLUSIONS

The reliability of the microprocessor-based TPSS has been studied in detail using conditional probability. Consideration of the four criteria; i.e., the probability of outage, the probability of activity, the improvement factor, and the activity factor, should provide an adequate description of the effectiveness of the overall protection plan. Several conclusions can be drawn from the analysis. First, terminal circuits usually have low failure rates so that one protection channel is adequate for the protection of many regular channels without having excessive

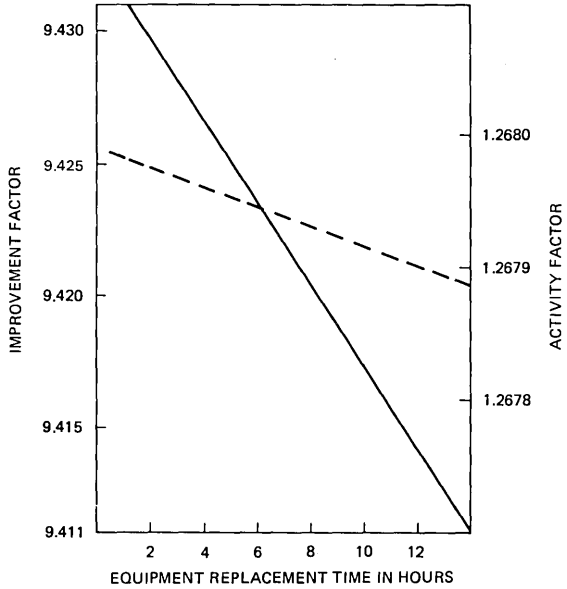


Fig. 15—Improvement and activity factors as functions of equipment replacement time.

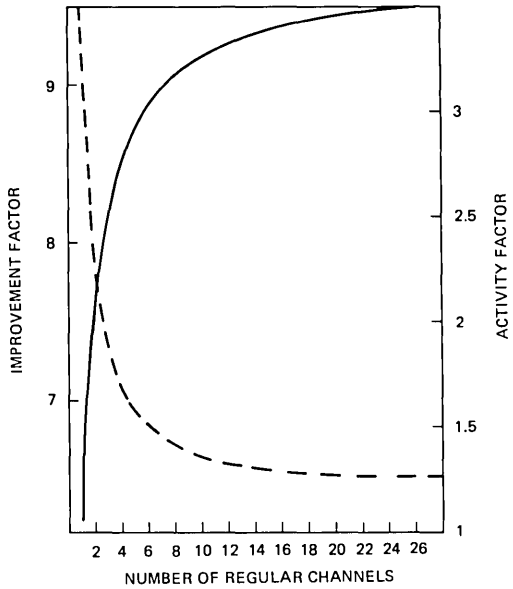


Fig. 16—Improvement and activity factors as functions of number of regular channels.

probability of service outage. Second, undetectable failures are usually the prime causes for increased outage probability and decreased improvement factor. If preventive maintenance is ever to be carried out, the hidden failures should be the principal targets. Third, the microcomputer is reliable as a protection switching controller. Although microprocessor system failures can cause false switching all by themselves, they contribute only a very small amount of the total outage if adequate self-checking is implemented. Reliability could be further improved by providing hardware interlock logic to guard against an insane microprocessor. For example, logic circuit can be provided in the TPSS to prevent the operation of an output switch whenever its input switch is inactive. Fourth, all the figures indicate that, around the various estimated failure rates of interest, the outage probabilities increase almost linearly with the failure rates. Thus there is no "preferred" range of failure rates that any equipment should be designed to. Only the sensitivities of the outage probabilities to the various estimates are different. Fifth, for any TPSS, the implementation of a 1×1 protection plan should be studied carefully. Even if there is improvement in the outage probability due to equipment failure, the increased activity will generate more maintenance-induced outages, not to mention increased costs.

The above comments do not apply in line protection switching systems, which have much higher regular channel failure rates because of the cascaded repeaters. Finally, Fig. 17 suggests one more consideration in determining the length of a line protection switching span. The failure

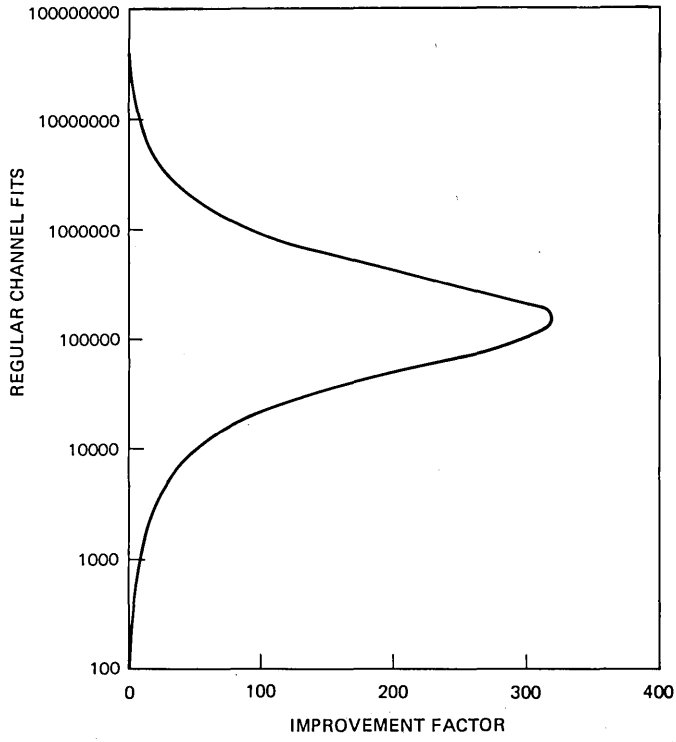


Fig. 17—Regular channel failure rates as functions of improvement factor.

rate of the line should preferably not fall into the decreasing region of its improvement factor. The last two points are obvious and interesting protection switching behavior patterns which seem not to have been explictly pointed out before.

APPENDIX A

This appendix discusses microprocessor self-test algorithms whose purpose is to generate alarms as early as possible to initiate maintenance actions. The test should be exhaustive but should not require too much additional program memory. An 8-bit microprocessor is used in the TPSS application.

When the power is turned on, the microprocessor immediately performs a thorough RAM check. Static RAMs are used, so there is no pattern sensitivity problem. The checking algorithm is to write the least-significant 8-address bits of each RAM byte into that specific RAM location. After all RAM locations are loaded, the contents of each byte are compared with its least-significant 8-bit address. After a byte is checked, its contents are complemented and checked again. The complemented contents will remain in those bytes already checked. This algorithm is

able to detect any bit, any data pin, and any combination of address pins stuck to zero or one. It can also discover data and address lines shorted together. Thus most RAM failures can be detected.

The ROMs are checked immediately following the RAM check. Two consecutive bytes in each ROM are reserved for self-test. One byte is used for parity check and the other for short-circuits in address and data lines. The microprocessor reads out every byte in the ROM and performs a cumulative odd parity check through an exclusive-OR operation on each bit. It will be seen first that, as far as independent ROM bit failures are concerned, it is adequate to use only one byte to check the parity of all ROMs no matter how many ROMs are used in the system. Let ℓ be the number of ROM bytes (excluding the reserved checking byte) used in the system and ϵ be the probability of a ROM bit failure. The probability of having parity violations is $1 - (1 - p)^8$, where p is[6]

$$\left[\frac{1 - (1 - 2\epsilon)^{\ell}}{2} \times (1 - \epsilon) + \frac{1 + (1 - 2\epsilon)^{\ell}}{2} \times \epsilon\right].$$

The probability of having bit errors is simply $1 - (1 - \epsilon)^{(\ell+1)\times 8}$. For $\ell\epsilon \ll 1$, both probabilities can be approximated by $8 \times (\ell + 1) \times \epsilon$. Thus the single byte parity check is adequate when $\ell\epsilon \ll 1$. It can be seen below that this condition is always valid in practice. Since the experimental failure rate of the 1K-byte EROM is 300 FITS, the failure rate of each bit cannot be more than $300/(8 \times 1024) \approx 0.037$ FIT. If a ROM failure can be discovered in 24 hours, then $\epsilon < 10^{-9}$. The number ℓ is limited by the microprocessor addressing capability which is 64K. Therefore, $\ell\epsilon \ll 1$. The reason that one parity byte is used in each ROM is to detect address and data lines stuck to one or zero. Since the ROM has a capacity equal to a power of 2, a stuck output looks like an even number of ones or zeros and violates the odd parity. A stuck address will cause half the bytes to be read twice and again violate the odd parity.

The contents of the bits of the other byte used for self-test are alternating ones and zeros. When this byte is read, short-circuits in data lines are detected. If this byte is located at an address whose 10 least-significant address bits are alternating ones and zeros, reading this byte will most likely detect short-circuits among these address lines. The probability is very small that within the same ROM another byte which also contains alternating ones and zeros is read because of shorted address lines. To detect some of the short-circuits in the remaining six most significant address lines, complemented numbers are stored in these checking bytes according to their address parities. Each ROM can select one of two hexidecimal numbers, AA or 55, to store at one of two addresses. For the first ROM with 0000 starting address, the two addresses are 0155 and 02AA.

The two consecutive checking bytes must be preceded by a jump or

branch instruction to bypass them in normal program execution. It is obvious that, if a single parity checking byte is located at an address with alternating ones and zeros, it alone can detect all ROM failures mentioned above except shorted data lines. It is sometimes possible to make use of the opcode and the operand of the jump or branch instruction to check the shorted data lines. If any failure occurs in the first ROM where the checking program is stored, the failure cannot always be detected. Duplicating the first ROM may be a possible solution.

After the two memory tests, a few instructions are exercised to test the CPU. Then the microprocessor starts executing the main program. Under normal circumstances, the program never comes back to the above RAM, ROM, and CPU tests. Different checks are performed in the main program. To avoid delaying the program execution, only distributed checks on the memory system are made. For example, in going through a program loop, only one RAM byte is tested and only one ROM exclusive OR is taken. However, the ROM check uses the same algorithm discussed above. The RAM check uses alternating ones and zeros which detect only shorted data lines and stuck bits because the exhaustive RAM check discussed before will destroy the temporary data stored, in addition to requiring long execution time. After each cycle of the nonexhaustive RAM check, an additional test[7] is made. Zeros are stored in the first RAM byte. Ones are stored only in RAM bytes with addresses 2^i, $i = 1,2,\cdots$. Every time all ones are loaded into an address, the contents of the first all-zero byte are also checked. The check is also distributed so as not to delay normal program execution. Most remaining RAM failures can be discovered by this additional test.

The effectiveness of the two RAM checking algorithms discussed above is similar. The first one used when turning on the power requires fewer steps and is faster. The second one does not destroy any temporary data because every check involves at most two RAM bytes (the first byte and the 2^ith byte) whose contents can be temporarily stored into CPU registers.

No CPU check is performed in the main program. A restarting sanity timer is employed to detect CPU failures. Under normal operation, the program retriggers the timer at durations shorter than the length of the timer. If the timer times out, an alarm is generated and the microprocessor system will go through its power on restart cycle again. The restarting sanity timer detects complete CPU failures. It can sometimes catch other CPU failures (for example, program counter skipping). It also reduces the damages that are caused by power transients because it restarts the system. RAM failures sometimes cause the timer to time out. ROM failures have similar effects but are more difficult to be self-detected. Output failures can only be detected by reading back the output bits immediately after each output operation.

APPENDIX B

This appendix derives the probabilities of outage and activity with and without protection switching. Figure 1 shows the configuration for a $1 \times n$ protection switching system in each direction of transmission. The microprocessor is responsible for the switching actions of $2n$ regular channels. The unprotected system has only the regular transmission channels plus pilot detectors for alarm.

The events of interests in deriving the outage probabilities are

S: service outage without protection switching.
S_P: service outage with protection switching.
G_1: all regular channels are good.
G_2: both protection channels are good.
G_3: all regular detectors are good.
G_4: all through switches are good.
G_5: all substitute switches are good.
G_6: the microprocessor system is good.
G_7: all output switches are good.

The events G_i's are assumed to be statistically independent. Their probabilities are given by

$$P\{G_1\} = p_r^{2n}$$
$$P\{G_2\} = p_p^2$$
$$P\{G_3\} = p_d^{2n}$$
$$P\{G_4\} = p_t^{2n}$$
$$P\{G_5\} = p_s^{2n}$$
$$P\{G_6\} = p_m = p_c p_e p_a$$
$$P\{G_7\} = p_0^{2n},$$

where the notations are defined in Table I. The symbol q with appropriate subscripts is defined to be $1 - p$ with the same subscript. Let $\overline{G}_i$ be the complement of G_i and g be the joint events of the G_i's with subscripts denoting the complemented events. For instance,

$$g_0 = G_1 G_2 G_3 G_4 G_5 G_6 G_7$$

and

$$g_{35} = G_1 G_2 \overline{G}_3 G_4 \overline{G}_5 G_6 G_7.$$

If these events represent all the possible failure modes of the system, then

$$P\{S_P\} = P\{S_P g_0\} + P\{S_P g_1\} + \cdots + P\{S_P g_6\} + P\{S_P g_7\} + P\{S_P g_{12}\} + \cdots + P\{S_P g_{234567}\} + P\{S_P g_{1234567}\}. \quad (1)$$

There are a total of 2^7 terms in (1). Half the terms involve the event $\overline{G}_7$, which generates service outage regardless of the other events. Therefore,

$$P\{S_P\} = 1 - p_0^{2n} + P\{S_P g_0\} + \cdots + P\{S_P g_6\} + P\{S_P g_{12}\} + \cdots + P\{S_P g_{23456}\} + P\{S_P g_{123456}\}. \quad (2)$$

The 2^6 unknown terms in (2) are to be evaluated. Since the derivations of each term are very similar, only the details in obtaining the more involved $P\{S_P g_{1345}\}$ and $P\{S_P g_{26}\}$ will be given. From the definition of conditional probability,

$$\begin{aligned} P\{S_P/g_{1345}\} &= P\{S_P/g_{1345}, \text{ three or more channel failures}\} \\ &\quad \cdot P\{\text{three or more channel failures}/g_{1345}\} \\ &\quad + P\{S_P/g_{1345}, \text{ two channel failures}\}P\{\text{two channel failures}/g_{1345}\} \\ &\quad + P\{S_P/g_{1345}, \text{ one channel failure}\}P\{\text{one channel failure}/g_{1345}\}. \end{aligned} \quad (3)$$

It is obvious that two protection channels cannot protect three failures; hence

$$P\{S_P/g_{1345}, \text{ three or more channel failures}\} = 1.$$

The joint event of three or more regular channel failures and $\overline{G}_1 G_2 \overline{G}_3 \overline{G}_4 \overline{G}_5 G_6 G_7$ has the conditional probability

$$P\{\text{three or more channel failures}/g_{1345}\} = \frac{[1 - p_r^{2n} - 2np_r^{2n-1}q_r - n(2n-1)p_r^{2(n-1)}q_r^2] \times p_p^2(1 - p_d^{2n})(1 - p_t^{2n})(1 - p_s^{2n})p_m p_0^{2n}}{P\{g_{1345}\}}. \quad (4)$$

The second term in (3) will be evaluated next. The various events will be abbreviated by their initials after their full names are introduced; e.g., tcf represents two channel failures.

$$\begin{aligned} P\{S_P/g_{1345}, \text{ tcf}\} &= P\{S_P/g_{1345}, \text{tcf, both failures in the same direction of transmission}\} \cdot P\{\text{both failures in the same direction of transmission}/g_{1345}, \text{tcf}\} \\ &\quad + P\{S_P/g_{1345}, \text{tcf, one failure in each direction}\} \cdot P\{\text{one failure in each direction}/g_{1345}, \text{tcf}\} \\ &= 1 \cdot \{n(n-1)p_r^{2(n-1)}q_r^2 p_p^2(1 - p_d^{2n})(1 - p_t^{2n})(1 - p_s^{2n})p_m p_0^{2n}\}/ \\ &\quad P\{g_{1345}, \text{tcf}\} + P\{S_P/g_{1345}, \text{tcf, one failure in each direction}\} \cdot P\{\text{ofied}/g_{1345}, \text{tcf}\}. \end{aligned} \quad (5)$$

Equation (5) follows because one protection channel cannot protect two failures in the same direction of transmission. The second term of (5) gives

$$P\{S_P/g_{1345}, \text{tcf,ofied}\} = P\{S_P/g_{1345}, \text{tcf,ofied, two associated detectors are not both good}\} \cdot P\{\text{two associated}$$

$$\text{detectors are not both good}/g_{1345},\text{tcf,ofied}\}$$
$$+ P\{S_P/g_{1345},\text{tcf,ofied, two associated detectors good}\}$$
$$\cdot P\{\text{two associated detectors good}/g_{1345},\text{tcf, ofied}\}$$
$$= 1 \cdot [n^2 p_r^{2n-2} q_r^2 p_p^2 (1 - p_d^2)(1 - p_t^{2n})(1 - p_s^{2n}) p_m p_0^{2n}]/$$
$$P\{g_{1345},\text{tcf,ofied}\} + P\{S_P/g_{1345},\text{tcf,ofied,tadg}\} \cdot P\{\text{tadg}/g_{1345},\text{tcf,ofied}\}. \quad (6)$$

$$P\{S_P/g_{1345},\text{tcf,ofied,tadg}\} = P\{S_P/g_{1345},\text{tcf,ofied,tadg,}$$
$$\text{both associated substitute switches good}\} \cdot P\{\text{both}$$
$$\text{associated substitute switches good}/g_{1345},\text{tcf,ofied,tadg}\}$$
$$+ 1 \cdot [n^2 p_r^{2n-2} q_r^2 p_p^2 p_d^2 (1 - p_d^{2n-2})(1 - p_t^{2n})(1 - p_s^2) p_m p_0^{2n}]/$$
$$P\{g_{1345},\text{tcf,ofied,tadg}\}. \quad (7)$$

$$P\{S_P/g_{1345},\text{tcf,ofied,tadg,bassg}\} = P\{S_P/g_{1345},\text{tcf,ofied,}$$
$$\text{tadg,bassg, both associated through switches good}\}$$
$$\cdot P\{\text{both associated through switches good}/g_{1345},\text{tcf,ofied,}$$
$$\text{tadg,bassg}\} + P\{S_P/g_{1345},\text{tcf,ofied,tadg,bassg, not both}$$
$$\text{through switches good}\} \cdot P\{\text{not both through switches}$$
$$\text{good}/g_{1345},\text{ttcf,ofied,tadg,bassg}\}$$
$$= 1 \cdot [n^2 p_r^{2n-2} q_r^2 p_p^2 p_d^2 (1 - p_d^{2n-2}) p_t^2 (1 - p_t^{2n-2}) p_s^2 (1 - p_s^{2n-2}) p_m p_0^{2n}]/$$
$$P\{g_{1345},\text{tcf,oefied,tadg,bassg}\} + P\{S_P/g_{1345},\text{tcf,ofied,tadg,}$$
$$\text{bassg, nbtsg}\} \cdot P\{\text{nbtsg}/g_{1345},\text{tcf,ofied,tadg,bassg}\}. \quad (8)$$

For the first term in (8), it is known that not all through switches are good because of $\overline{G}_4$. The outage probability is one because if the two failed channels have good through switches, the rest of the through switches must have failure. Finally,

$$P\{S_P/g_{1345},\text{tcf,ofied,tadg,bassg,nbtsg}\} = P\{S_P/g_{1345},\text{tcf,}$$
$$\text{ofied,tadg,bassg,nbtsg, no other through switch failure}\}$$
$$\cdot P\{\text{no other switch failure}/g_{1345},\text{tcf,ofied,tadg,}$$
$$\text{bassg,nbtsg}\} + P\{S_P/g_{1345},\text{tcf,ofied,tadg,bassg,nbtsg, other}$$
$$\text{through switch failure}\} \cdot P\{\text{other through switch}$$
$$\text{failure}/g_{1345},\text{tcf,ofied,tadg,bassg,nbtsg}\}$$
$$= 0 + [n^2 p_r^{2n-2} q_r^2 p_p^2 p_d^2 (1 - p_d^{2n-2})(1 - p_t^2)(1 - p_t^{2n-2})$$
$$\cdot p_s^2 (1 - p_s^{2n-2}) p_m p_0^{2n}]/P\{g_{1345},\text{tcf,ofied,tadg,bassg,nbtsg}\}. \quad (9)$$

In (9), the first conditional outage probability is zero because all the failures are protected by the two protection channels. The above derivations illustrate one of the basic approaches. Each event and its complement are assumed until the conditional probability of outage is either one or zero.

The third term in (3) is similarly derived.

$$P\{S_P/g_{1345},\text{ocf}\} = P\{S_P/g_{1345},\text{ocf, associated detector bad}\}$$

$$\cdot P\{\text{associated detector bad}/g_{1345},\text{ocf}\} + P\{S_P/g_{1345},\text{ocf, associated detector good}\} \cdot P\{\text{associated detector good}/g_{1345},\text{ocf}\}$$
$$= 1 \cdot [2np_r^{2n-1}q_r p_p^2 q_d(1-p_t^{2n})(1-p_s^{2n})p_m p_0^{2n}]/P\{g_{1345},\text{ocf}\} + P\{S_P/g_{1345},\text{ocf,adg}\}P\{\text{adg}/g_{1345},\text{ocf}\}. \quad (10)$$

$$P\{S_P/g_{1345},\text{ocf,adg}\} = P\{S_P/g_{1345},\text{ocf,adg, associated substitute switch good}\} \cdot P\{\text{associated substitute switch good}/g_{1345},\text{ocf,adg}\} + 1 \cdot [2np_r^{2n-1}q_r p_p^2 p_d(1-p_d^{2n-1}) \times (1-p_t^{2n})q_s p_m p_0^{2n}]/P\{g_{1345},\text{ocf,adg}\}. \quad (11)$$

$$P\{S_P/g_{1345},\text{ocf,adg,assg}\} = P\{S_P/g_{1345},\text{ocf,adg,assg, one other through switch bad}\} \cdot P\{\text{one other through switch bad}/g_{1345},\text{ocf,adg,assg}\} + 1 \cdot \{2np_r^{2n-1}q_r p_p^2 p_d(1-p_d^{2n-1})[1-p_t^{2n-1}-(2n-1)p_t^{2n-2}q_t] \cdot p_s(1-p_s^{2n-1})p_m p_0^{2n}\}/P\{g_{1345},\text{ocf,adg,assg}\}. \quad (12)$$

Equation (12) indicates that the status of the through switch associated with the failed regular channel has no effect on the outage probability.

$$P\{S_P/g_{1345},\text{ocf,adg,assg,ootsb}\} = P\{S_P/g_{1345},\text{ocf,adg,assg,ootsb, bad through switch in other direction of transmission}\} \cdot P\{\text{bad through switch in other direction}/g_{1345},\text{ocf,adg,assg,ootsb}\} + 1 \cdot [2np_r^{2n-1}q_r p_p^2 p_d(1-p_d^{2n-1})p_t^n(n-1)p_t^{n-2}q_t p_s \times (1-p_s^{2n-1})p_m p_0^{2n}]/P\{g_{1345},\text{ocf,adg,assg,ootsb}\}. \quad (13)$$

$$P\{S_P/g_{1345},\text{ocf,adg,assg,ootsb,btsiod}\} = P\{S_P/g_{1345},\text{ocf,adg,assg,ootsb,btsiod, bad switch has good detector}\} \cdot P\{\text{bad switch has good detector}/g_{1345},\text{ocf,adg,assg,ootsb,btsiod}\} + 1 \cdot [2np_r^{2n-1}q_r p_p^2 p_d q_d n p_t^{2n-2}q_t p_s(1-p_s^{2n-1})p_m p_0^{2n}]/P\{g_{1345},\text{ocf,adg,assg,ootsb,btsiod}\}. \quad (14)$$

$$P\{S_P/g_{1345},\text{ocf,adg,assg,ootsb,btsiod,bshgd}\} = P\{S_P/g_{1345},\text{ocf,adg,assg,ootsb,btsiod,bshgd, corresponding substitute switch bad}\} \cdot P\{\text{corresponding substitute switch bad}/g_{1345},\text{ocf,adg,assg,ootsb,btsiod,bshgd}\} + 0 \cdot P\{\text{corresponding substitute switch good}/g_{1345},\text{ocf, adg,assg,ootsb,btsiod,bshgd}\}$$
$$= 1 \cdot [2np_r^{2n-1}q_r p_p^2 p_d^2(1-p_d^{2n-2})np_t^{2n-2}q_t p_s q_s p_m p_0^{2n}]/P\{g_{1345},\text{ocf,adg,assg,ootsb,btsiod,bshgd}\}. \quad (15)$$

From (3) through (15),

$$P\{S_P g_{1345}\} = p_p^2 p_m p_0^{2n}\{(x+x_3)(1-p_d^{2n})(1-p_t^{2n})(1-p_s^{2n})$$

$$
\begin{aligned}
&+ x_1[q_d(1-p_t^{2n})(1-p_s^{2n}) + p_d(1-p_d^{2n-1})(1-p_t^{2n})q_s \\
&+ p_d(1-p_d^{2n-1})\cdot[1-p_t^{2n-1}-(2n-1)p_t^{2n-2}q_t]p_s(1-p_s^{2n-1}) \\
&+ p_d(1-p_d^{2n-1})\cdot(n-1)p_t^{2n-2}q_tp_s(1-p_s^{2n-1}) \\
&+ p_dq_dnp_t^{2n-2}q_tp_s(1-p_s^{2n-1}) + p_d^2(1-p_d^{2n-2})np_t^{2n-2}q_tp_sq_s] \\
&+ x_4[(1-p_d^2)(1-p_t^{2n})(1-p_s^{2n}) + p_d^2(1-p_d^{2n-2})(1-p_t^{2n})(1-p_s^2) \\
&+ p_d^2(1-p_d^{2n-2})p_t^2(1-p_t^{2n-2})p_s^2(1-p_s^{2n-2}) + p_d^2(1-p_d^{2n-2}) \\
&\qquad\cdot(1-p_t^2)(1-p_t^{2n-2})p_s^2(1-p_s^{2n-2})]\}, \quad (16)
\end{aligned}
$$

where

$$
\begin{aligned}
x_1 &= 2np_r^{2n-1}q_r \\
x_2 &= 1-p_r^{2n}-2np_r^{2n-1}q_r \\
x_3 &= 1-p_r^{2n}-2np_r^{2n-1}q_r-n(2n-1)p_r^{2n-2}q_r^2 \\
x_4 &= n^2p_r^{2n-2}q_r^2 \\
x &= n(n-1)p_r^{2n-2}q_r^2.
\end{aligned}
$$

To evaluate $P\{S_Pg_{26}\}$, the events

H_1: CPU is good
H_2: ROMs are good
H_3: RAMs are good

will be considered separately. Let h represent joint events similar to those for g, for example, $h_2 = H_1\overline{H}_2H_3$. As before,

$$
\begin{aligned}
P\{S_P/g_{26}\} &= P\{S_P/g_{26}, \text{ both protection channels bad}\}P\{\text{both protection channels bad}/g_{26}\} + P\{S_P/g_{26}, \text{ one protection channel bad}\}P\{\text{one protection channel bad}/g_{26}\}P\{S_P/g_{26},\text{bpcb}\} \\
&= P\{S_P/g_{26},\text{bpcb},h_1\}P\{h_1/g_{26},\text{bpcb}\} + P\{S_P/g_{26},\text{bpcb},h_2\} \\
&\quad\times P\{h_2/g_{26},\text{bpcb}\} + P\{S_P/g_{26},\text{bpcb},h_3\}P\{h_3/g_{26},\text{bpcb}\} \\
&\quad + P\{S_P/g_{26},\text{bpcb},h_{12}\}P\{h_{12}/g_{26},\text{bpcb}\} + P\{S_P/g_{26},\text{bpcb},h_{13}\} \\
&\quad\times P\{h_{13}/g_{26},\text{bpcb}\} + P\{S_P/g_{26},\text{bpcb},h_{23}\}P\{h_{23}/g_{26},\text{bpcb}\} \\
&\quad + P\{S_P/g_{26},\text{bpcb},h_{123}\}P\{h_{123}/g_{26},\text{bpcb}\}. \quad (17)
\end{aligned}
$$

The microprocessor operation is so complicated that simplifying assumptions have to be made before (17) can be further evaluated. There are two kinds of CPU failures. The first kind is a partial failure which may not be detected by the self-checking method discussed in Appendix A. For instances, program counter skipping and one CPU transistor failure within the CPU may not always be detectable. This partial failure may generate false switching and result in service outage. The second kind is a complete failure, and the CPU operation stops altogether. No false switching will be made in this case, and the sanity timer will detect the

failure immediately. It is assumed that partial failures accounts for 20 percent of the total CPU failures.

When the CPU is partially failed, it executes the contents of the ROMs insanely. Every "instruction" has a finite probability of generating a false switching. The TPSS software contains approximately 4000 bytes of which 100 can be I/O instructions. Out of the $2n + 5$ hardware addresses, $2n$ have outputs controlling the switches. If a correct parity bit and an appropriate output switch control bit are stored in the accumulator, an I/O instruction will operate the output switch. If the protection channels are bad, the operation of the output switch will generate service outage regardless of the status of the input switch. Thus the probability p_1 that any instruction will cause an outage is approximately

$$p_1 = \frac{100}{4000} \cdot \frac{1}{4} \cdot \frac{2n}{2n+5}.$$

When the protection channels are working, the same probability is now

$$p_2 = \frac{100}{4000} \cdot \frac{1}{8} \cdot \frac{2n}{2n+5}$$

because the input switch should be inactive for the false output switching to generate service outage. It is to be noted that false switching can also occur randomly if the 8-bit "instruction," the 16-bit "address," the parity bit, and the switch control bit happen to match the real instruction and address. This probability is of the order $2n/2^{26}$ and is negligible compared with p_1 and p_2. On the average, each instruction takes about 4 microseconds. Thus before restoration, about

$$n_1 = \frac{\mu_c \times 60 \times 60 \times 10^6}{4}$$

"instructions" are executed. The probability p_3 that an outage will occur is

$$\begin{aligned} p_3 &= p_1 + q_1 p_1 + \cdots + q_1^{n_1-1} p_1 \\ &= p_1 \frac{1 - q_1^{n_1}}{1 - q_1} \\ &= 1 - q_1^{n_1}. \end{aligned}$$

When the protection channels are good, the corresponding probability is

$$p_4 = 1 - q_2^{n_1}.$$

After a false switching, it is possible that insane CPU may deactivate the switch and restore service. It may also operate other output switches to generate additional service outages. These two conditional proba-

bilities are small. If they are ignored, the outage probability assuming partial CPU failure and bad protection channels is then $p_3 t/\mu_c$. If only one of the two protection channels is bad, let

$$p_5 = \frac{100}{4000} \cdot \frac{1}{8} \cdot \frac{n}{2n+5} + \frac{100}{4000} \cdot \frac{1}{4} \cdot \frac{n}{2n+5}.$$

The outage probability is $p_6 t/\mu_c$ where

$$p_6 = 1 - q_5^{n1}.$$

When a memory failure occurs, the program counter jumps to an arbitrary location. The initial effect is somewhat like that of a partially failed CPU. Experiments indicate that outage is unlikely to occur if it has not occurred during the initial period. Since 25 out of the 4000 bytes are used to activate the output switches in normal program operation, a jump to these bytes will cause a false switching. Therefore, the false switching probability is

$$p_7 = \frac{25}{4000} + p_1$$

or

$$p_8 = \frac{25}{4000} + p_2,$$

depending on whether the protection channels are bad or good. If only one of the two protection channels is bad, the probability is

$$p_9 = \frac{25}{4000} + p_5.$$

It will be assumed that all RAM failures can be detected. Most of the RAM bytes are used for stack. The effects of the ROM and the RAM failures are assumed to be identical, but their restoration times are different because not all ROM failures are self-detectable. When the CPU fails, memory failures are assumed to have no effect on the system. This makes the evaluation of the fourth, the fifth, and the last terms in (17) unnecessary once the first term is evaluated. It is further assumed that when there are both ROM and RAM failures, the trouble can be detected immediately. Given the previous assumption, then

$$\begin{aligned} P\{S_P/g_{26},\text{bpcb},h_1\} &= P\{S_P/g_{26},\text{bpcb},h_1,\text{ complete failure}\}P\{\text{complete failure}/g_{26},\text{bpcb},h_1\} \\ &\quad + P\{S_P/g_{26},\text{bpcb},h_1,\text{ partial failure}\}P\{\text{partial failure}/g_{26},\text{bpcb},h_1\} \\ &= 0 + \frac{p_3 t}{\mu_c} \frac{p_{10} q_p^2 0.2 q_c p_e p_a}{P\{g_{26},\text{bpcb},h_1\}}, \end{aligned}$$

where

$$p_{10} = (p_r p_d p_t p_s p_0)^{2n}. \tag{18}$$

$$P\{S_P/g_{26},\text{bpcb},h_2\} = \frac{p_7 t}{\mu_e}\frac{p_{10}q_p^2 p_c q_e p_a}{P\{g_{26},\text{bpcb},h_2\}}$$

$$P\{S_P/g_{26},\text{bpcb},h_3\} = p_7\frac{p_{10}q_p^2 p_c p_e q_a}{P\{g_{26},\text{bpcb},h_3\}}$$

$$P\{S_P/g_{26},\text{bpcb},h_{23}\} = p_7\frac{p_{10}q_p^2 p_c q_e q_a}{P\{g_{26},\text{bpcb},h_{23}\}}.$$

Hence

$$P\{S_P,g_{26},\text{bpcb}\} = p_{10}\left\{\frac{p_3 t}{\mu_c}0.2\,q_c + \frac{p_7 t}{\mu_e}p_c q_e p_a + p_7 p_c q_a\right\}. \quad (19)$$

The expression $P\{S_P,g_{26},\text{opcb}\}$ can be similarly evaluated. Finally,

$$P\{S_P g_{26}\} = p_{10}\left\{2p_p q_p\left[\frac{p_6 t}{\mu_c}0.2q_c + \frac{p_9 t}{\mu_e}p_c q_e p_a + p_9 p_c q_a\right] + q_p^2\left[\frac{p_3 t}{\mu_c}0.2q_c + \frac{p_7 t}{\mu_e}p_c q_e p_a + p_7 p_c q_a\right]\right\}. \quad (20)$$

After deriving (16) and (20), the remaining terms in (2) are easy to obtain. They will not be given here. Thus the outage probability with protection switching $P\{S_p\}$ is obtained from (2). It should be emphasized that, because there are hidden failures, multiple equipment failures cannot be neglected in evaluating the various terms in (2). In fact, the term that contributes the most to the outage probability is $P\{S_p g_{135}\}$, which involves both of the undetectable failures (detector and substitute switch).

Since the detectors used to generate alarms do not affect signal transmission, the outage probability without protection switching is simply

$$P\{S\} = 1 - p_r^{2n}. \quad (21)$$

The improvement factor is

$$\text{IF} = \frac{P\{S\}}{P\{S_P\}}. \quad (22)$$

Next, the probabilities of activity with and without protection switching will be considered. The additional events of interest are

A: activity without protection switching
A_P: activity with protection switching
G_5: protection detectors are good.

G_5 is redefined because protection detector failures generates maintenance activities, but the hidden substitute switch failures are assumed to cause no activity. To calculate the probability of activity with protection switching, notice that whenever $\overline{G}_1$, $\overline{G}_4$, and $\overline{G}_7$ occur, there will definitely be maintenance activity. Furthermore, the events $\overline{G}_2$ and $\overline{G}_5$ are detectable when G_6 is true. Therefore

$$P\{A_P\} = 1 - (p_r p_t p_0)^{2n} + (p_r p_t p_0)^{2n} p_m (1 - p_p^2 p_D^2) + P\{A_P g_0\} + P\{A_P g_3\} + P\{A_P g_6\} + P\{A_P g_{26}\} + P\{A_P g_{36}\} + P\{A_P g_{56}\} + P\{A_P g_{236}\} + P\{A_P g_{256}\} + P\{A_P g_{356}\} + P\{A_P g_{2356}\}. \quad (23)$$

In (23), $P\{A_P g_0\}$ is always zero. The last seven terms are negligible compared with $P\{A_P g_3\}$ and $P\{A_P g_6\}$. It is assumed that 10 percent of the CPU and the ROM failures will not generate alarm. The derivation of $P\{A_P g_6\}$ is similar to that of (17). For example,

$$P\{A_P/g_6 h_1\} = P\{A_P/g_6, h_1, \text{ undetectable failure}\} P\{\text{undetectable failure}/g_6 h_1\} + P\{A_P/g_6, h_1, \text{ detectable failure}\} P\{\text{detectable failure}/g_6, h_1\}$$

$$= 0 + \frac{t}{\mu_c} \cdot (p_r p_d p_t p_0)^{2n} (p_p p_D)^2 \cdot 0.9 \cdot q_c p_e p_a / P\{g_6 h_1\}.$$

Thus,

$$P\{A_P g_6\} = (p_r p_d p_t p_0)^{2n} (p_p p_D)^2 \left[0.9 \frac{t}{\mu_c} q_c + 0.9 \frac{t}{\mu_e} p_c q_e p_a + p_c q_a \right]. \quad (24)$$

If it is assumed that, when a detector fails, the probability that it is stuck to an ON state is 0.25, then

$$P\{A_P/g_3\} = P\{A_P/g_3, \text{ one detector bad}\} P\{\text{one detector bad}/g_3\} + \cdots + P\{A_P/g_3, 2n \text{ detectors bad}\} P\{2n \text{ detectors bad}\}. \quad (25)$$

The ith term in (25) is

$$P\{A_P/g_3, \text{idb}\} = P\{A_P/g_3, \text{idb, all bad detectors on}\} P\{\text{all bad detectors on}/g_3, \text{idb}\} + P\{A_P/g_3, \text{idb, some bad detectors off}\} \cdot P\{\text{some bad detectors off}/g_3, \text{idb}\}$$

$$= 0 + \frac{t}{\mu_d} \cdot \frac{(p_r p_t p_0)^{2n} (p_p p_D)^2 p_m \binom{2n}{i} p_d^{2n-i} q_d^i (1 - 0.25^i)}{P\{g_3, \text{idb}\}}.$$

Therefore,

$$P\{A_P g_3\} = \frac{t}{\mu_d} p_m (p_p p_D)^2 (p_r p_t p_0)^{2n} \sum_{i=1}^{2n} p_d^{2n-i} q_d^i (1 - 0.25^i). \quad (26)$$

Equations (23) through (26) yield the probability of activity with protection switching $P\{A_P\}$. The probability of activity without protection switching $P\{A\}$ is simply

$$P\{A\} = 1 - p_r^{2n} + \frac{t}{\mu_b} p_r^{2n} \sum_{i=1}^{2n} \binom{2n}{i} p_b^{2n-i} q_b^i (1 - 0.25^i),$$

where

$$p_b = \frac{1}{1 + \lambda_d \mu_b}$$

and

$$\mu_b = \frac{1}{4\lambda_r}$$

is the detector restoration time without protection switching. The activity factor is given by

$$\text{AF} = \frac{P\{A_P\}}{P\{A\}}. \quad (27)$$

REFERENCES

1. "L5 Coaxial-Carrier Transmission System," B.S.T.J., *53,* No. 10 (December 1974), pp. 1897–2268.
2. J. A. Buzacott, "Markov Approach to Finding Failure Times of Repairable Systems," IEEE Trans. on Reliability, *R-19* (November 1970), pp. 128–134.
3. B. V. Gnedenko, Y. K. Belyayev, and A. D. Solovyev, *Mathematical Methods of Reliability Theory* (Russian orig. and English transl. edited by R. E. Barlow), New York: Academic Press, 1969.
4. I. Welber, H. W. Evans, and G. A. Pullis, "Protection of Service in the TD-2 Radio Relay System by Automatic Channel Switching," B.S.T.J., *34,* No. 3 (May 1955), pp. 473–510.
5. W. Y.-S. Chen, "Estimated Outage in Long-Haul Radio Delay Systems With Protection Switching," B.S.T.J., *50,* No. 4 (April 1971), pp. 1455–1485.
6. G. S. Fang, "Alarm Statistics of the Violation Monitor and Remover," B.S.T.J., *55,* No. 8 (October 1976), pp 1197–1217.
7. B. A. Zimmer, "Test Techniques for Circuit Boards Containing Large Memories and Microprocessor," IEEE Computer Society Conf. Proceedings, Semiconductor Test Symposium, Oct. 19 to 21, 1976, Cherry Hill, N. J.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 7, September 1978
Printed in U.S.A.

Offset Multireflector Antennas with Perfect Pattern Symmetry and Polarization Discrimination

By C. DRAGONE

(Manuscript received September 20, 1977)

Conditions are derived that are useful for designing reflector antennas with excellent cross-polarization discrimination. These conditions ensure circular symmetry and absence of cross-polarization everywhere in the far field of an antenna, provided a suitable feed such as a corrugated horn is employed. The spherical wave radiated by the fundamental mode of such a feed has circular symmetry around the axis, and it is everywhere free of cross-polarization. An arbitrary sequence of N confocal reflectors (hyperboloids, ellipsoids, paraboloids) is combined with such a feed. It is shown that it is always possible to ensure circular symmetry (and absence of cross-polarization) in the antenna far field by properly choosing the feed axis orientation. If the final reflector is a paraboloid, a simple geometrical procedure can be used. It is also shown that the asymmetry caused by an arbitrary number of reflections can always be eliminated by properly introducing an additional reflection. An application to the problem of producing a horizontal beam using a vertical feed is discussed. Two arrangements are described that may be useful for radio relay systems.

Use of orthogonal polarizations is often required in radio systems to double transmission capacity. Antennas providing good discrimination between the two polarizations are then needed. The main purpose of this paper is to derive and discuss certain conditions that ensure excellent discrimination. When two or more reflectors and a suitable feed are arranged in accordance with these conditions, the antenna far field has, in all directions, the same polarization of the feed excitation. Furthermore, its pattern has circular symmetry. The above conditions also minimize astigmatism, and for this reason they are also useful* in the design of multibeam antennas (with several feeds).

* This is the subject of an article being prepared.

I. INTRODUCTION

A suitable feed for the antennas considered here is realized by properly corrugating the walls of a circular horn.[1–4] The spherical wave radiated by the horn then has circular symmetry and, by placing the feed at the focus of a paraboloid, an antenna with circular symmetry in the far field is obtained, provided the paraboloid is centered around the feed axis. Furthermore, the polarization of the plane wave reflected by the paraboloid then coincides with that of the feed excitation.

However, in the centered configuration the reflected wave is in part blocked by the horn.* To avoid this, the horn axis can be offset as in Fig. 1, but unfortunately this causes asymmetry in the pattern after reflection, resulting in an undesired cross-polarized component.[5,6] The same behavior occurs if, instead of a paraboloid, an arbitrary reflector system with a single axis of revolution is used. In Fig. 1, the asymmetry of the reflected wave increases with the angle of incidence α of the ray corresponding to the horn axis. This particular ray will be called *principal ray*.

Although a single offset reflection always causes some asymmetry, it

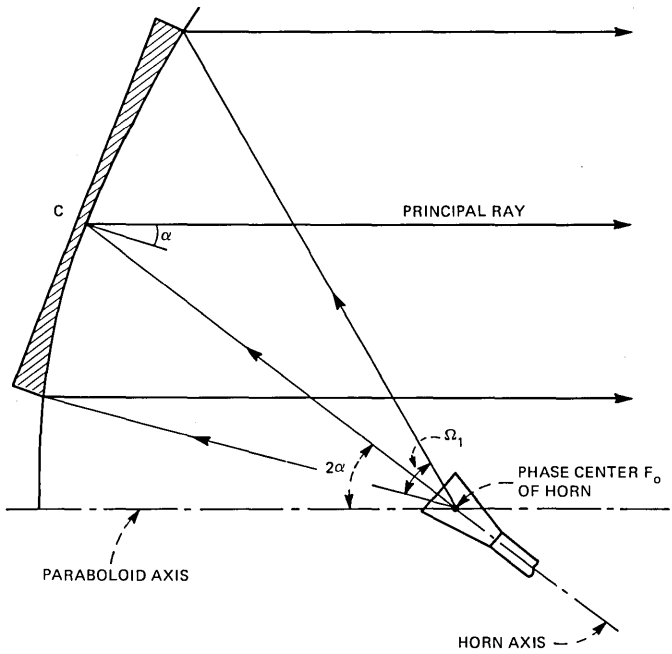


Fig. 1—The spherical wave radiated from F_0 by a corrugated feed is transformed by an offset paraboloid into a plane wave.

* This blockage impairs gain, side-lobes level, return loss, and cross-polarization discrimination.

is possible to combine two reflections with nonzero angles of incidence so as to ensure perfect symmetry after the two reflections.[7–10] In this paper we generalize and extend the results of Refs. 7 to 9 in several respects. First, the analysis here is not restricted to only two reflections, nor does it assume the final reflector is necessarily a paraboloid. Second, very simple conditions that guarantee symmetry after the final reflection are obtained. These conditions are shown to be direct consequences of a general principle of equivalence (see the appendix). Third, a general solution is given to the problem* of restoring the symmetry of a wave whose initial symmetry has been distorted by an arbitrary number of reflectors.

In Section III, two arrangements with excellent performance in cross-polarization are described. Both arrangements produce a horizontal beam using a vertical feed and may therefore be useful for microwave radio systems.

The following analysis is based on geometrical optics. Furthermore, the far field for the antennas of Figs. 12 and 13 is not derived in Section III, but it is important to note that the principle of equivalence of the following section allows the aperture field distribution for both antennas to be derived replacing the reflectors with a single paraboloid, centered around the feed axis. The aperture field distribution and far field of such a paraboloid are well known.[1–5] As pointed out at the beginning of this introduction, the entire aperture will be polarized in one direction if the feed is linearly polarized. The far field is thus free of cross-polarization, neglecting secondary effects such as edge diffraction.

II. THE EQUIVALENT REFLECTOR AND THE ORIENTATION OF ITS AXIS

Suppose a spherical wave from F_0, initially with symmetrical pattern, is successively reflected N times, using paraboloids, hyperboloids, and ellipsoids as shown in Fig. 2 for $N = 3$. The reflectors are properly arranged so that a spherical wave is produced after each reflection. Thus, if F_n is the focal point after the nth reflection, the nth reflector Σ_n transforms a spherical wave centered at F_{n-1} into a spherical wave centered at F_n. Note that some of the points $F_0, F_1, \cdots, F_N$ may be at ∞, in which case the corresponding spherical waves become plane waves. In Fig. 2, F_3 is at ∞, and therefore the last reflector is a paraboloid.

It is shown in the appendix that *such a sequence of confocal reflectors is always equivalent to a single reflector* which will be either an ellipsoid, a hyperboloid, or a paraboloid. This equivalent reflector produces, after a single reflection, the same reflected wave† as the given sequence of

* An interesting formulation of this problem is given in Ref. 10.

† Thus, if one considers the field distribution over a wavefront reflected by the equivalent reflector, it will coincide with the field distribution over the corresponding wavefront produced by the given sequence of reflectors.

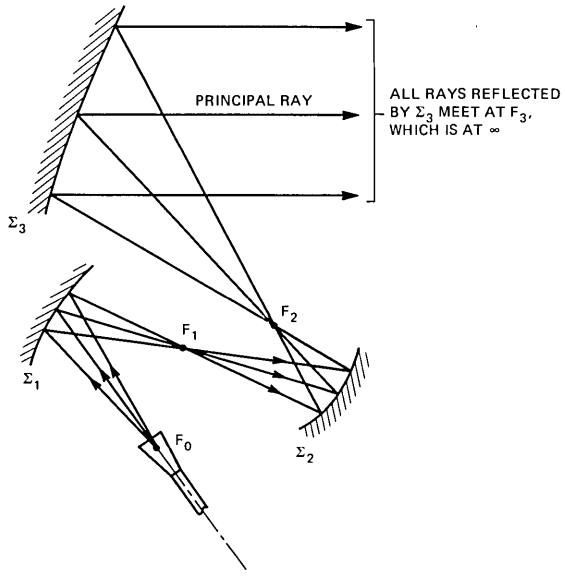


Fig. 2—The spherical wave from F_0 is transformed into a plane wave by three confocal reflectors. The nth reflector transforms the spherical wave from F_{n-1} into a spherical wave converging towards F_n.

reflectors. Thus, for the purpose of determining the properties of the reflected wave, one may replace the N reflectors with the equivalent reflector. This reflector has an axis of symmetry, which passes through F_0, and will be called the *equivalent axis.* It is clear that in order that the symmetry of the incident beam be preserved, *the principal ray must coincide with the equivalent axis.**

2.1 The central rays, their closed path, and the equivalent axis

Consider first $N = 1$. Suppose the reflector Σ_1 and one of its foci, F_0, are given, but the exact location of the axis of Σ_1 is not known and must be found. Then one may proceed as follows. Let a ray from F_0 be reflected twice by Σ_1, as shown in Fig. 3, and let $\vec{s}$ and $\vec{s}''$ be the initial and final directions of the ray. Then, from Fig. 3,

$$\vec{s} = \vec{s}'' \tag{1}$$

only when the ray coincides with the axis. Thus, the axis can be found by searching for a ray that satisfies this condition. Note from Fig. 3 there are two such rays, with opposite directions.

Next consider $N > 1$. Since a confocal sequence of reflectors $\Sigma_1, \cdots, \Sigma_N$ is equivalent to a single reflector Σ_e, the above procedure is appli-

* Since one can travel along the equivalent axis in two opposite directions, two opposite orientations can be chosen for the principal ray.

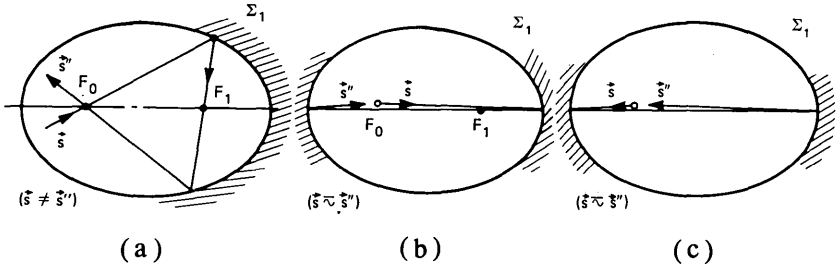


Fig. 3—The axis of Σ_1 is determined by varying $\vec{s}$ until $\vec{s} = \vec{s}''$.

cable also to this case. Thus, to determine the axis of Σ_e (equivalent axis), one must consider a ray from F_0, with initial directions $\vec{s}$. This ray must be reflected *twice* by Σ_e, and $\vec{s}$ must then be chosen so that $\vec{s}'' = \vec{s}$. Notice that the two reflections by Σ_e imply a total of $2N$ reflections in the original configuration. The first N reflections take place in the order $\Sigma_1, \cdots, \Sigma_N$, while the last N have the reverse order $\Sigma_N, \cdots, \Sigma_1$. The final ray passes again through F_0, with the same direction as the original ray. In Fig. 4a, $\vec{s} \neq \vec{s}''$. In Fig. 4b, on the other hand, condition (1) is satisfied, and therefore the ray through F_0 gives the correct orientation of the equivalent axis (and the principal ray for which symmetry is preserved).

Notice that if, after the above $2N$ reflections, the ray in Fig. 4a is reflected $2N$ more times it will not follow the same path of the first $2N$ reflections. In Fig. 4b, on the other hand, the path of the first $2N$ reflections is closed. This closed path, which determines the equivalent axis, will be called the *central path*. The two rays that proceed along the central path in opposite senses will be called the *central rays*.

We show next that condition (1) leads to a straightforward geometrical procedure for determining the equivalent axis when Σ_N is a paraboloid.

*2.2 The equivalent axis when the last reflector Σ_N is a concave paraboloid**

It is now shown that, when the last ellipsoid in Fig. 4a is replaced by a concave paraboloid, the final ray direction $\vec{s}''$ becomes independent of the initial direction $\vec{s}'$. This constant value of $\vec{s}''$ then gives the direction of the equivalent axis, which can thus be found straightforwardly.

Notice the path of Fig. 4a involves two successive reflections by the last ellipsoid Σ_N. Let ψ be the angle between the axis of Σ_N and the ray produced after the second reflection (see Fig. 5). The parameters of the ellipsoid Σ_N are now gradually modified, keeping the vertex V and the focus F_{N-1} fixed, increasing the distance between F_N and F_{N-1} until

* The following considerations apply also when Σ_N is a convex paraboloid, but this case is of little practical interest and will therefore be ignored.

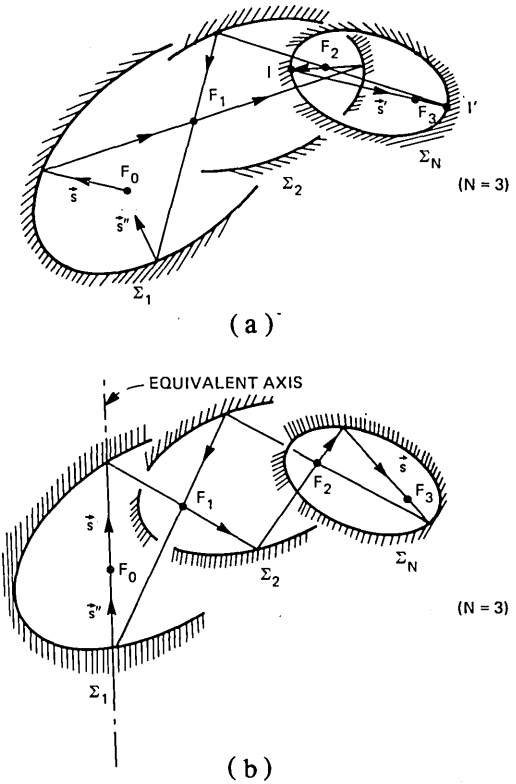


Fig. 4—(a) $2N$ successive reflections. (b) The central path. The equivalent axis through F_0 is obtained by varying in (a) the initial direction $\vec{s}$ until $\vec{s} = \vec{s}''$ as shown in (b).

$F_N \to \infty$. The ellipsoid then becomes a paraboloid with focus F_{N-1} and from the figure $\psi = 0$, which shows that

> If a ray from the focus F_{N-1} of a paraboloid is reflected twice by the paraboloid, so that the second reflection occurs at ∞, the final ray coincides with the paraboloid axis and it has the direction going from F_{N-1} towards the vertex V of the paraboloid. (2)

This implies that, when in Fig. 4 the last ellipsoid Σ_N is replaced by a paraboloid, the direction of $\vec{s}''$ becomes independent of $\vec{s}$, and it can be determined by tracing the ray $F_{N-1}V$ as shown in Fig. 6. The direction $\vec{s}''$ so obtained gives the equivalent axis, as one may verify considering a ray with *initial* direction given by the above value of $\vec{s}''$. One can see from Fig. 6 the path of this ray closes, after $2N$ reflections. Thus,

To obtain the equivalent axis of a sequence of $N - 1$ re-

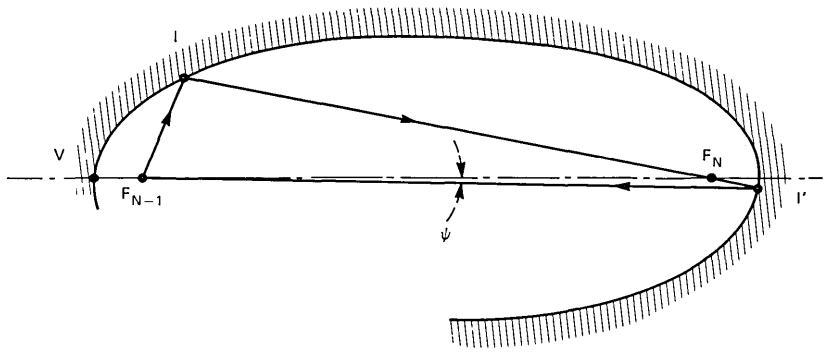


Fig. 5—As the distance of F_N from the other focus F_{N-1} is increased, keeping V and F_{N-1} fixed, the ellipsoid approaches a paraboloid with vertex V and focus F_{N-1}; for the ray reflected at I' one has $\psi \to 0$.

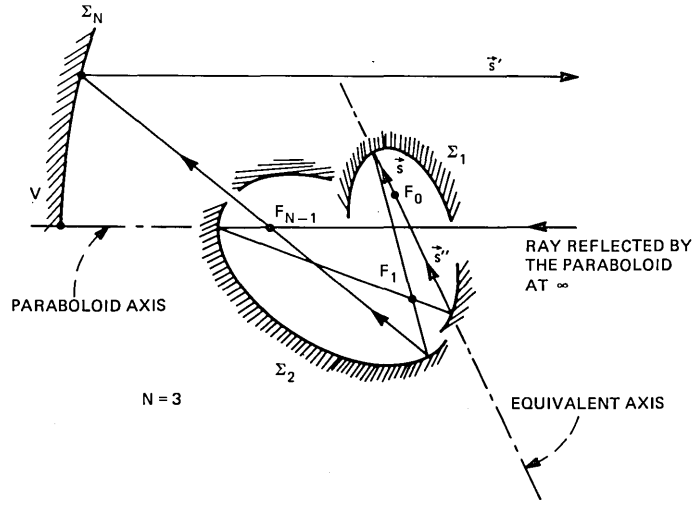


Fig. 6—By tracing from ∞ the path of the ray defined by the paraboloid axis one obtains after $N-1$ reflections the equivalent axis through F_0. If a symmetrical feed is placed at F_0, centered around the equivalent axis, a symmetrical pattern will be reflected by the paraboloid.

flectors $\Sigma_1, \Sigma_2, \cdots, \Sigma_{N-1}$ followed by a paraboloid Σ_N with focus F_{N-1} and vertex V, simply reflect $N-1$ times the ray $F_{N-1}V$ by $\Sigma_{N-1}, \Sigma_{N-2}, \cdots, \Sigma_1$. The final ray through F_0 is the equivalent axis and, therefore, the principal ray along which symmetry is preserved. (3)

As an example, consider $N = 2$, and assume the first reflector is not

a paraboloid.* Then four different arrangements are obtained depending on whether the first reflector is an ellipsoid or an hyperboloid, and is convex or concave. In each case (see Figs. 7 and 8), the equivalent axis† is determined by the intersection I' of the paraboloid axis with the first reflector. The equivalent axis is the line F_0I'. Note the axis of the paraboloid intercepts the first reflector Σ_1 in two points, but only one, I', is acceptable.‡ The acceptable point is the point of reflection of the ray F_1V. Since only one side of the surface Σ_1 is reflecting, only one of the above two points can be considered a point of reflection for the above ray.

From Figs. 7 and 8, since in all cases the equivalent axis and the paraboloid axis meet on Σ_1, the angles 2α and 2β giving their inclinations from the axis of Σ_1 are related,

$$\tan \alpha = m \tan \beta, \tag{4}$$

where m is the axial magnification of Σ_1 given by the distances of the reflector vertex V_0 from the two focal points F_0 and F_1,

$$m = \frac{|F_0V_0|}{|F_1V_0|}. \tag{5}$$

Note that if e is the eccentricity of the reflector, in Figs. 7 and 8,

$$m = \frac{e+1}{e-1}, \frac{e-1}{e+1}, \frac{e+1}{1-e}, \frac{1-e}{1+e}, \tag{6}$$

respectively. In Fig. 7 one has $e > 1$, whereas in Fig. 8, $0 < e < 1$.

In the two cases of Figs. 7a and 8a, eq. (4) is equivalent to eq. (1) of Ref. 9. In the other two cases, on the other hand, eq. (1) of Ref. 9 is not applicable [to obtain a correct relation, one has to replace α with β in eq. (1)].

Another useful relation, derived in the following section, is

$$\tan i = \frac{M}{1-M} \tan p. \tag{7}$$

It relates the angles of incidence i and p of the central ray on the two

* The case where Σ_1 is a paraboloid is treated in Section 2.6.

† That is, the beam orientation for which symmetry is preserved.

‡ Notice for the purpose of deriving the equivalent axis that the entire surfaces of the various ellipsoids, hyperboloids, and paraboloids must be considered to be reflecting. Thus, both branches of an hyperboloid must be considered. Of course, an actual antenna will use only certain sections of the various surfaces.

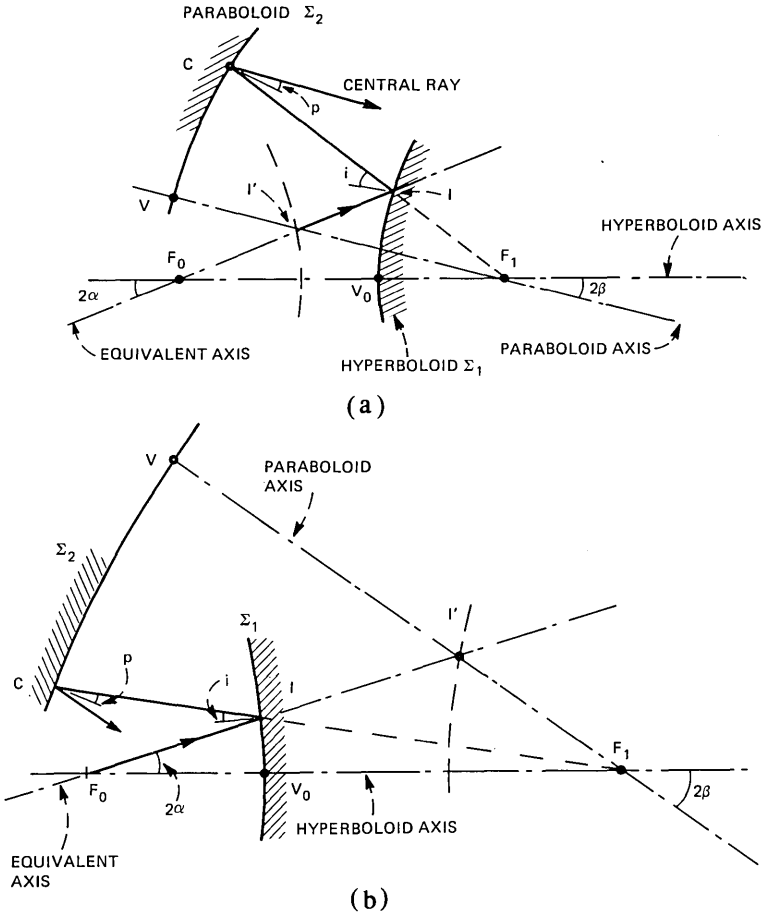


Fig. 7—How to determine the central path and the equivalent axis of a paraboloid combined in (a) with a convex hyperboloid and in (b) with a concave hyperboloid.

reflectors (see Figs. 7a and 8) to the magnification M, defined as

$$M = \pm \frac{|F_0 I|}{|I F_1|}, \tag{8}$$

I being the point of incidence of the central ray on the first reflector. In eq. (8) one has to take the positive sign when F_0 and F_1 are on opposite sides of the tangent plane at I, as in Fig. 8; otherwise, as in Fig. 8, $M <$ 0. The angles of incidence must be taken with opposite sign in Figs. 7a and 8, where the two reflections have opposite senses; in Fig. 7b, on the other hand, i and p have the same sign.

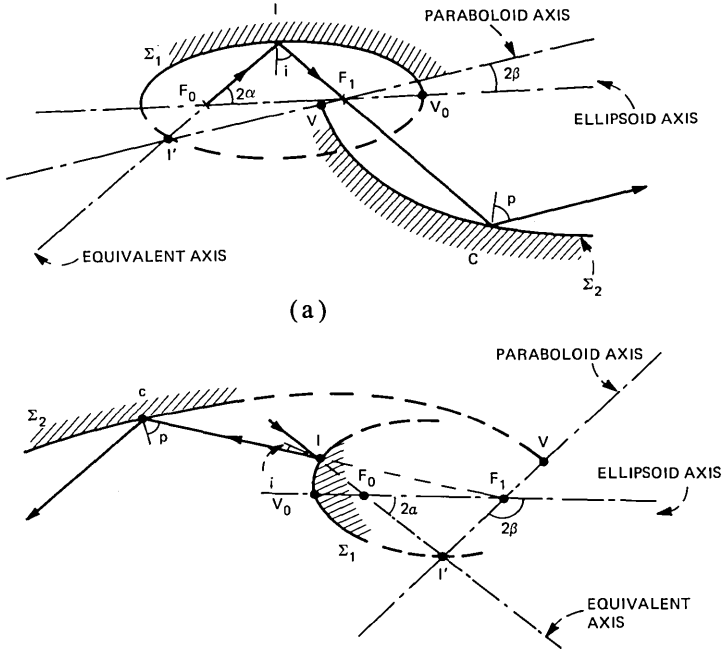


Fig. 8—How to determine the central path and the equivalent axis of a paraboloid combined in (a) with a concave ellipsoid and in (b), with a convex ellipsoid.

The magnification* M determines the ratio between the angular width Ω_0 of the beam incident as I and the width Ω_1 of the reflected beam. More precisely,† for small Ω_0,

$$M = \frac{\Omega_1}{\Omega_0}. \tag{9}$$

If M is specified, eq. (7) gives the angles of incidence i and p that result in a symmetrical beam after two reflections.

A very general relation, which reduces to eq. (7) in the particular case where Σ_N is a paraboloid, is derived in Section 2.4.

---

* Another important significance of M is that the paraxial focal length f_e, for any of the arrangements of Figs. 7 and 8, in the vicinity of the central ray, is $f_e = Mf_p$, where f_p is the paraboloid focal length $f_p = CF_1$; f_e has the significance that a small lateral displacement δs of a feed initially placed at F_0 will cause an angular displacement $\delta\theta = \delta s/f_e$ of the beam reflected by the paraboloid.

† Thus, if a beam of small angular width Ω_0 is transformed by a sequence of N reflectors with magnifications $M_1, \cdots, M_N$, the final beam has angular width

$$\Omega_1 = M_t\Omega_0,$$

where $M_t = M_1M_2 \cdots M_N$.

2.3 Relations governing the reflections of a central ray by the first or the last reflector

The restriction that Σ_N must be a paraboloid is now removed. The closed path of the central ray in Fig. 4 involves two successive reflections by Σ_1. Consider these two reflections and assume for the moment Σ_1 is a concave ellipsoid as shown in Fig. 9a. The central ray in Fig. 9a first passes through F_1 with direction $\vec{a}$, it is successively reflected at I' and I, and it then passes again through F_1 with direction $\vec{c}$.

Let $2i$ and $2i'$ be the angles of the two reflections and M and M' the corresponding magnifications,

$$M = -\frac{\ell_1}{\ell_2}, M' = -\frac{\ell_1'}{\ell_2'}. \tag{10}$$

ℓ_1, ℓ_2, etc. being defined in Fig. 9a. Then, if $2\gamma = 2i + 2i'$ is the total angle of reflection (given by the angle between the final and initial rays $\vec{c}$ and $\vec{a}$) it is shown in Section A.3 of the appendix that

$$\tan i = \frac{M}{M-1} \tan \gamma \tag{11}$$

and

$$\tan i' = \frac{1}{1-M'} \tan \gamma. \tag{12}$$

Thus, if the parameters (M,i, or M',i') of either reflection are given, the total angle of reflection for a central ray can be calculated. Note that eqs. (11) and (12) apply also to the two consecutive reflections of the central ray by the last reflector Σ_N.

In Fig. 9a, the reflector Σ_1 is a concave ellipsoid, but eqs. (11) and (12) are valid also if Σ_1 is an hyperboloid or is concave, as shown in Figs. 9b, c, and d. Note in cases 9c and 9d the central ray is first reflected at I', ,then passes through the point at ∞ and is then reflected again at I. Figs. 7a,b and 8a,b correspond to Figs. 9b, 9c, 9a, and 9d, respectively.

2.4 How to arrange two reflectors

Consider Fig. 10a showing a principal ray from F_0 reflected by two reflectors Σ_1 and Σ_2. We wish to show that, in order that this ray be a central ray, i.e., that symmetry be preserved after these two reflections, their parameters M, M', i, and i' must satisfy the condition

$$\tan i = M \frac{1-M'}{1-M} \tan i'. \tag{13}$$

Consider the ray reflected by Σ_1. Let this ray be reflected *twice* by Σ_2, and then again *twice* by Σ_1, as in Fig. 10b. If 2γ denotes the total angle of the first two reflections by Σ_2 and $2\gamma'$ the angle of the other two re-

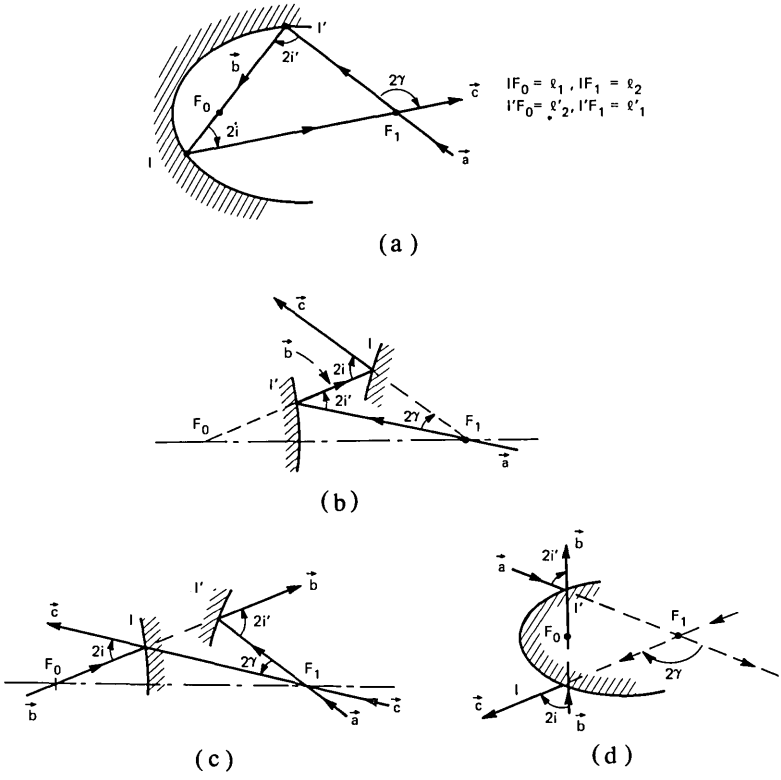


Fig. 9—Two successive reflections. (a) By concave ellipsoid. (b) By convex hyperboloid. (c) By concave hyperboloid. (d) By convex ellipsoid.

flections, one must have

$$2\gamma + 2\gamma' = 2\pi, \tag{14}$$

if the path of the ray is to close (which is necessary for it to be a central ray) after the four consecutive reflections. Now $\tan\gamma$ is given by eq. (11), and $\tan\gamma'$ by eq. (12) with γ replaced by γ'. Thus, by requiring condition (14), one obtains condition (13). In the particular case where the second reflector is a paraboloid,

$$M' = 0$$

and eq. (13) give Eq. 7 (with $i' = p$).

2.5 Restoration of beam symmetry after an arbitrary number of reflections

Suppose an arbitrary sequence of $N-1$ reflections $\Sigma_1, \cdots \Sigma_{N-1}$ have distorted the initial symmetry of a spherical wave originating from F_0. We wish to restore symmetry by introducing an additional reflector Σ_N. Let the principal ray through F_0 be reflected $N-1$ times by the given reflectors as shown in Fig. 11a for $N = 3$. The reflector Σ_N must be chosen

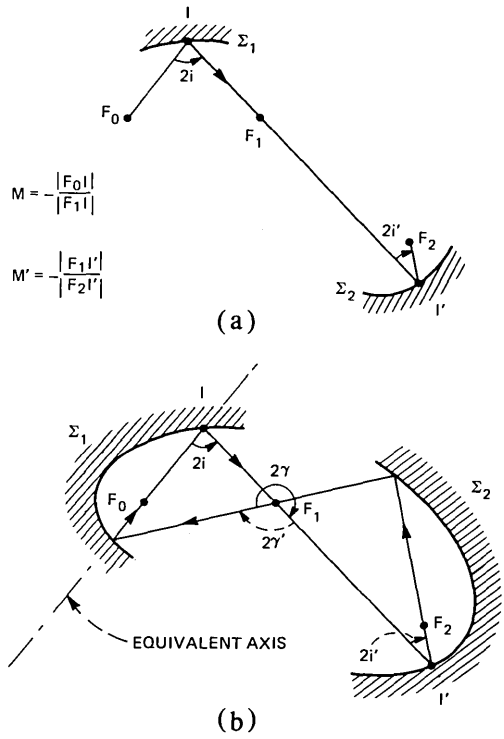


Fig. 10—Central path and equivalent axis of a combination of two ellipsoids.

so that this ray is one of the two *central rays* of the sequence $\Sigma_1, \cdots, \Sigma_N$. This means the path of the ray must *close* after $2N$ successive reflections. Now a part of this path, the section determined by the reflections of Σ_1, $\Sigma_2, \cdots, \Sigma_{N-1}$, is fixed in advance. Therefore let this part of the central ray be determined first. It starts at F_{N-1} and, after $2(N-1)$ reflections, it ends again at F_{N-1} with direction $\vec{a}$ as shown in Fig. 11a. Since its final direction $\vec{a}$ is given, its initial direction $\vec{c}$ can be found by tracing the ray backwards. Once $\vec{c}$ is known, the condition that Σ_N must satisfy is simply eq. (12), with γ given by the angle between $\vec{c}$ and $\vec{a}$, shown in Fig. 11.

2.6 How to determine the first reflector if the remaining ones are given

The above argument applies also to the problem where the first reflector, rather than the last, is to be found and the remaining reflectors are given. The only difference in this case is that one must use eq. (11), instead of eq. (12), as shown by the following example. To consider a situation of practical interest, suppose the last reflector Σ_N is a paraboloid as shown in Fig. 11b. Assume that all the reflectors except the first one are given and that Σ_1 must be chosen so that the central ray passes

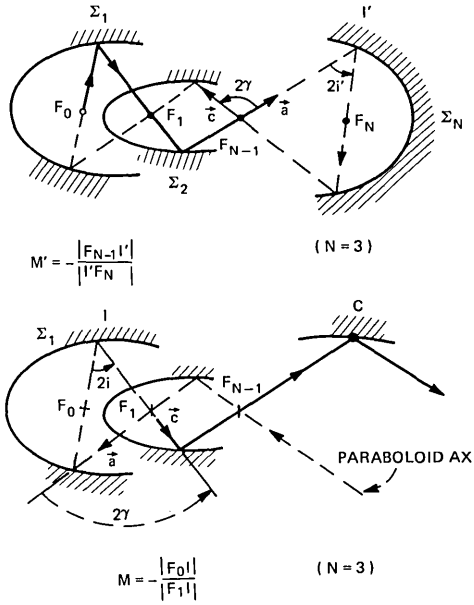


Fig. 11—(a) How to determine the last reflection if the first $N-1$ are given. (b) How to determine the first reflection if the last $N-1$ are given.

through the center C of the paraboloid aperture. Then, as in the previous problem, one notices that a part of the desired central path is fixed in advance. This part starts as F_1 with direction $\vec{c}$ and, after $2(N-1)$ reflections, it ends at F_1 with direction $\vec{a}$ as shown in Fig. 11b. Once $\vec{a}$ is found (by ray tracing), the condition that Σ_1 must satisfy is given by eq. (11), with γ given by the angle shown in the figure between $\vec{c}$ and $\vec{a}$.

2.7 The first and the last reflector are paraboloids

Consider first $N = 2$, in which case eq. (13) with $M = M' = \infty$ demands that the angles of incidence on the two paraboloids be identical, except for a difference in sign. For this to happen, the axes of the two paraboloids must coincide, in which case one can show that the two angles of incidence coincide *for any* choice of the principal ray. These remarks apply also to $N > 2$, since the last $N-1$ reflectors can always be replaced by an equivalent paraboloid. Thus,

> In order that symmetry be preserved, when both Σ_1 and Σ_N are paraboloids, the axis of Σ_1 must coincide with the equivalent axis of $\Sigma_2, \cdots, \Sigma_N$, in which case symmetry is preserved by any choice* of the principal ray. (15)

* A little thought shows that there is another case where the central ray is undetermined: namely, when the equivalent reflector is a flat plate.

III. AN APPLICATION

The most important example of an offset arrangement is perhaps the horn reflector,[11] an antenna consisting of a horn combined with a paraboloid. The excellent properties of this antenna (negligible return loss, very low level of the far sidelobes, etc.) are well known. However, the angle of incidence on the paraboloid is 45 degrees, and this causes in the far field a cross-polarized component of about −20 dB in certain directions.[11] The 45-degree angle of incidence is required to produce a beam orthogonal to the feed axis, which is an important requirement* for radio relay systems. In this section it is shown, with two examples given in Figs. 12 and 13, how this requirement can be fulfilled using two or more reflectors satisfying condition (7). In both Figs. 12 and 13, the feed is of the type described in Refs. 1 to 4, and therefore the antenna beam is essentially free of cross-polarization *everywhere* (see the last remark in the introduction).

Figure 12 shows two large reflectors, a paraboloid and an hyperboloid, arranged to satisfy simultaneously condition (7) and the requirement $i + p = 90°$, without aperture blockage. This arrangement is of the type shown in Fig. 8b of Ref. 7. In Fig. 13, three reflectors, a large paraboloid Σ_3, and two small hyperboloids Σ_2 and Σ_1 are used. This arrangement is more compact, and it requires less total reflecting area, than the one of Fig. 12. It is thus particularly attractive when the antenna aperture is large, i.e., the far-field beamwidth is small. The angle of incidence i and the magnification M of the first reflector Σ_1 satisfy condition (7), with p given by the angle shown in Fig. 12. To understand the significance of p, replace the last two reflectors Σ_2 and Σ_3 by their equivalent paraboloid. According to (3), the axis of this paraboloid is obtained from the axis of Σ_3 by reflecting it once, onto Σ_2, as shown in Fig. 13. Then $2p$ is the angle the central ray makes with this equivalent axis. Note that p is equal to the angle of incidence on this equivalent paraboloid (not shown in Fig. 13). This angle of incidence must satisfy eq. (7). One can verify from the figure that

$$\tan p = \frac{\tan \alpha + m_2 \tan \beta}{1 - m_2 \tan \alpha \tan \beta}, \tag{16}$$

α and β being the angles (see Fig. 7a) of the central ray and the axis of Σ_3 with respect to the axis of Σ_2, and

$$m_2 = \frac{|V_2F_1|}{|V_2F_2|} = \frac{e_2 + 1}{e_2 - 1}, \tag{17}$$

e_2 being the eccentricity of the hyperboloid Σ_2. Also,

$$2i = 90° + 2\beta - 2\alpha, \tag{18}$$

* Of course, this is not the only requirement that must be satisfied. Other requirements will be discussed in an article being prepared.

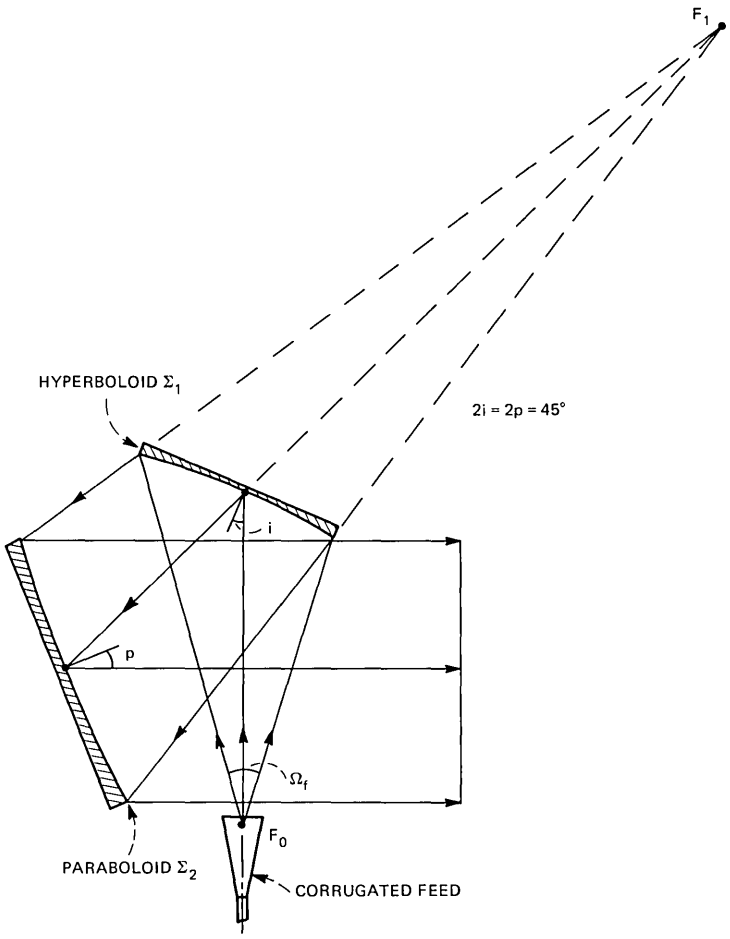


Fig. 12—A vertical feed and two reflectors with $i + p = 45$ degrees producing a horizontal beam without symmetry distortion.

and from eq. (7), solving for M,

$$M = \frac{\tan i}{\tan i + \tan p}. \tag{19}$$

Using eqs. (16) to (19), one can express M directly in terms of α, β, m_2.

An important property of Figs. 12 and 13 is that there is no aperture blockage even for relatively large values (as large as 30 degrees) of the angular width Ω_f of the beam radiated by the feed. Another important property, to be discussed in a future article, is that, if the feed is slightly displaced so as to cause a small angular displacement of the antenna beam, the resulting aberrations are very small. This is a consequence of

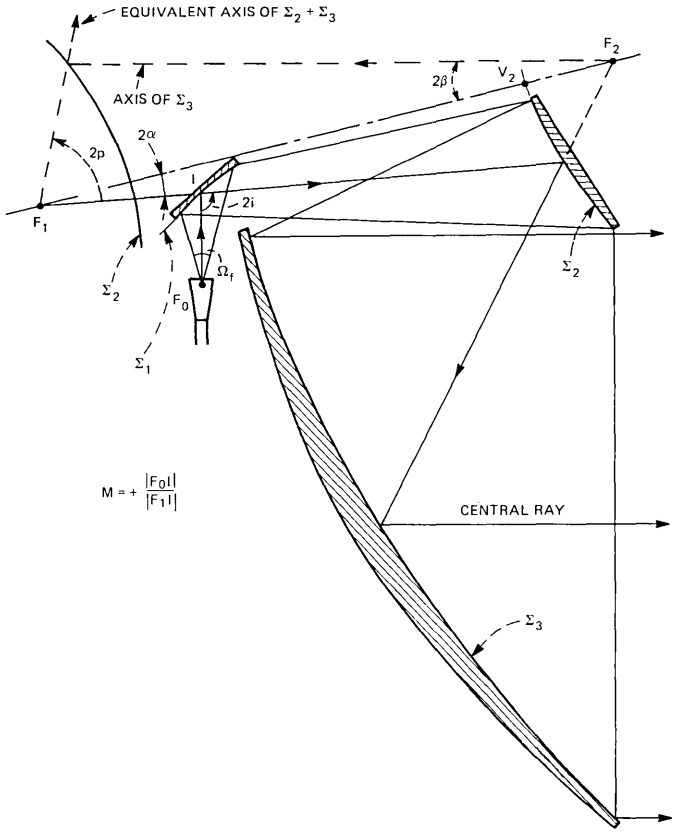


Fig. 13—A vertical feed and three reflectors producing a horizontal beam without symmetry distortion.

condition (7), and it implies that several beams can be produced efficiently by placing several feeds in the focal plane.

IV. CONCLUSIONS

The transformation of a symmetrical beam by an arbitrary arrangement of N confocal reflectors has been studied. It has been shown that it is always possible to choose the principal ray (i.e., the axis of the input beam) so that symmetry is preserved by the transformation. This is a consequence of the principle of equivalence shown in the appendix, according to which an arrangement of several reflectors can always be replaced by a single reflector producing the same transformation. Thus, in order that symmetry be preserved, the principal ray must coincide with the axis of symmetry of this equivalent reflector, i.e., the equivalent axis. A property of the equivalent axis is that the path of a ray having initially its direction becomes closed after $2N$ successive reflections. Because of this property, the equivalent axis can be found by a

straightforward geometrical procedure if the last reflector is a paraboloid. A simple relation [eq. (11) or (12)] has been given for determining the angle of incidence and the magnification of the first or last reflector so as to guarantee symmetry. In Section III, the problem of modifying the horn reflector to eliminate the asymmetry and cross-polarization due to the paraboloid has been discussed. Two solutions have been described.

APPENDIX

General Properties of a Sequence of N Confocal Reflectors

The results of this paper are consequences of the principle of equivalence stated at the beginning of Section II. This principle is now derived.

As pointed out in the introduction, the reflectors we consider are ellipsoids, hyperboloids, or paraboloids; let $F_0, F_1, \cdots, F_N$ be $N + 1$ arbitrary points, let a point source be placed at F_0, and let a sequence of N reflectors $\Sigma_1, \cdots, \Sigma_N$ be used to successively transform the spherical wave from F_0 into spherical waves through $F_0, F_1, \cdots, F_N$. The nth reflector, Σ_n, with its focal points of F_{n-1} and F_n then transforms the spherical wave incident from F_{n-1} into a spherical wave through F_n.

Draw two spheres S and S' centered at F_0 and F_N. For each point P of S, there is, on S', a corresponding point determined by the ray through P. This mapping has the following properties.

A circle on S′ corresponds to each circle on S. In fact, it is well known[12,13] that a circular cone of rays from F_{n-1} is transformed by the nth reflector into a circular cone of rays through F_n.

The mapping is conformal,* and therefore two orthogonal curves of S are transformed into two orthogonal curves of S'.

Another property is that, if the point source at F_0 is linearly polarized and the lines of the electric field $\vec{E}$ on S are given, then the corresponding lines defined on S' by the above mapping give correctly the lines of $\vec{E}$ on S'. This result is true in general[14] for arbitrary reflectors, not necessarily paraboloids, hyperboloids, or ellipsoids. It allows the polarization of S' to be determined straightforwardly once the relationship between corresponding rays through F_N and F_0 is known.

A.1 The central rays

Draw a line through F_0, to cut the sphere S at two antipodal points. We show that it is always possible to choose the line orientation so that the corresponding points of S' are also antipodal points.

* This property is valid in general for an arbitrary wavefront S which is transformed by an arbitrary number of reflections (by arbitrary reflectors, not necessarily of the type considered here) into a wavefront S'. The mapping determined between S and S' by the rays orthogonal to S (and S') is a conformal mapping.

Let L_1, L_2 and M_1, M_2 be antipodal points of S (see Fig. 14; the sphere S is not shown). Let L'_1, L'_2 and M'_1, M'_2 be their corresponding points on S'. Through $L'_1, L'_2\ M'_1, M'_2$ draw two great circles. The two circles will intersect in two antipodal points $0'_1$ and $0'_2$, as shown in Fig. 14. We show that the corresponding points are also antipodal points. In fact, 0_1 and 0_2 are the points of intersection of the two circles of S that correspond to the two circles of S'. Since the circles of S contain the antipodal points L_1, L_2 and M_1, M_2, they are great circles and therefore their intersections 0_1 and 0_2 are antipodal points. Q.E.D.

An important significance of the points 0_1, 0_2, $0'_1$, and $0'_2$ is the following. Let a ray from F_0 be reflected by the sequence of N reflectors *twice,* first in the order $\Sigma_1, \Sigma_2, \cdots, \Sigma_N$ and then in the reverse order Σ_N, $\Sigma_{N-1}, \cdots, \Sigma_1$. After these $2N$ reflections, the ray will pass again through F_0, but its direction will in general differ from the direction given initially, and therefore the ray will not in general follow the same path if reflected $2N$ more times. However, a little thought shows that, since the three points $0_1, F_0, 0_2$ are collinear and so also are $0'_1F_N0'_2$, the path of a ray from 0_1 (or from 0_2) will become closed after $2N$ reflections. The same observation applies to the ray from 0_2, which will follow, in the opposite direction, the same path of the ray from 0_1.

The path of the rays from 0_1 and 0_2 will be called the *central path* and the two rays *central rays.* This definition is consistent with the one given in Section II. As we shall see, there is in general only one central path, except when both Σ_1 and Σ_N are paraboloids (see Section 2.6) or when the equivalent reflector is a flat plate [$m_e = 1$ in eq. (21)].

The axial ray F_00_1 is now chosen as reference axis. Let a particular plane through this ray be chosen as reference plane. Consider a particular ray from F_0, and let θ be its angle with respect to the axis and ϕ the angle its plane makes with the reference plane. After N reflections, both the ray in question and the axial ray pass through F_N. Let θ' be the angle between the two rays at F_N, let ϕ' be the angle their plane makes with an arbitrary reference plane (chosen through the axial ray). We wish to

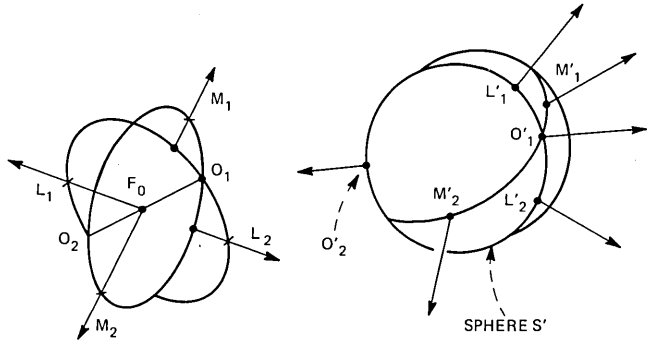


Fig. 14—How to determine the central rays.

show that

$$\phi' = \pm\phi + \phi_0 \tag{20}$$

and

$$\tan\frac{\theta'}{2} = m_e \tan\frac{\theta}{2}, \tag{21}$$

where m_e is a constant determined by the N reflectors and ϕ_0 is determined by the orientation of the two reference planes which will be chosen so that

$$\phi_0 = 0. \tag{22}$$

A.2 Derivation of eqs. (20) and (21)

First consider on S a great circle, through the two axial points 0_1 and 0_2, given by

$$\phi = a, \tag{23}$$

where a is a constant. Since the corresponding circle on S' must pass through $0_1'$ and $0_2'$, it is a great circle, given by

$$\phi' = a', \tag{24}$$

where a' is a constant. This shows that ϕ' depends only on ϕ, not on θ. We now recall that the mapping of S' must be conformal and therefore the angle between two circles through $0_1'$ must equal the angle between the corresponding circles of S. This implies eq. (20).

Next we derive eq. (21). Since the sign in front of ϕ in eq. (20) depends on the definition of ϕ', and can therefore be chosen arbitrarily, we choose for the following derivation

$$\phi' = \phi.$$

Since a circle θ = constant is orthogonal to a circle ϕ = constant, the corresponding circles on S' must be orthogonal. This implies θ' is a function of θ only. To determine this function, consider on S three points of coordinates:

$$(\theta,\phi), \quad (\theta + d\theta,\phi), \quad (\theta,\phi + d\phi).$$

Let

$$(\theta',\phi), \quad (\theta' + d\theta',\phi)(\theta',\phi + d\phi)$$

be the corresponding coordinates on S. Let $d\ell_1$ and $d\ell_2$ denote on S the distances of the first point from the other two. Then

$$d\ell_1 = rd\theta, \quad d\ell_2 = r\sin\theta d\phi, \tag{25}$$

r being the radius of the sphere S. Similarly, for the corresponding distances on S',

$$d\ell_1' = \mathrm{r}'d\theta', \quad d\ell_2' = \mathrm{r}'\sin\theta' d\phi. \tag{26}$$

Since the mapping is conformal, one must have

$$\frac{d\ell_1'}{d\ell_2'} = \frac{d\ell_1}{d\ell_2},$$

which gives the condition

$$\frac{d\theta}{\sin\theta} = \frac{d\theta'}{\sin\theta'}. \tag{27}$$

Integrating this gives eq. (21), where m_e is a constant of integration.

When $N = 1$, eqs. (20) and (21) are nothing new. In fact, then the reflector system reduces to a single reflector whose eccentricity determines the parameter m_e. When $N > 1$, eqs. (20) and (21) show the N reflectors are equivalent to a single reflector with eccentricity specified* by m_e.

A.3 Derivation of eqs. (11) and (12)

Consider the ellipsoid shown in Fig. 15. Then

$$\tan\alpha \tan\alpha' = 1 \tag{28}$$

and

$$\tan\alpha' \tan\psi' = \tan\alpha \tan\psi = m,$$

where

$$m = \frac{|F_0V_0|}{|F_1V_0|}. \tag{29}$$

Therefore, taking into account that $\gamma = 90° - \psi - \psi'$,

$$\tan\gamma = \frac{1 - \tan^2\psi \tan^2\alpha}{\tan\psi(1 + \tan^2\alpha)} \tag{30}$$

Also, $i = 90° - \alpha - \psi$, and therefore

$$\tan i = \frac{1 - \tan\alpha \tan\psi}{\tan\alpha + \tan\psi}. \tag{31}$$

Now the magnification M of I is defined as

$$M = -\frac{|IF_0|}{|IF_1|}, \tag{32}$$

and from Fig. 15 is related to the angles ψ and α,

$$M = -\frac{\sin 2\psi}{\sin 2\alpha}$$

$$= -\frac{\tan\psi}{\tan\alpha}\frac{1 + \tan^2\alpha}{1 + \tan^2\psi}, \tag{33}$$

* The value of m_e can be calculated using the formula

$$m_e = \pm M_1 \cdot M_2 \cdots M_N,$$

where $M_1, \cdots, M_N$ are the magnifications calculated for the N reflections of the central ray chosen as reference axis. The sign of m_e depends on the sign convention for ϕ in eq. (20).

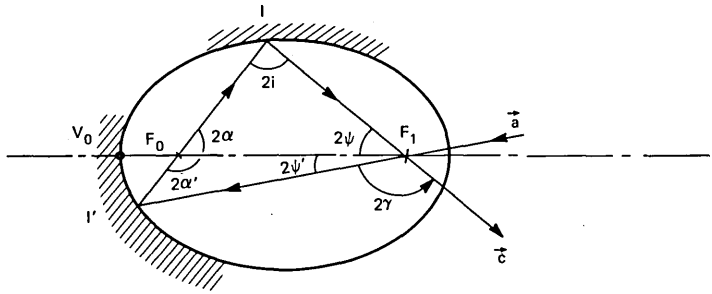


Fig. 15—Two successive reflections by a concave ellipsoid.

which gives

$$\frac{M}{M-1} = \frac{\tan\psi(1+\tan^2\alpha)}{(\tan\psi+\tan\alpha)(1+\tan\psi\tan\alpha)}. \tag{34}$$

From eqs. (30), (31), and (34), one obtains eq. (11). The derivation of eq. (12) is entirely analogous. The case where the reflector is convex, or is a hyperboloid, can be treated in the same way.

REFERENCES

1. V. H. Rumsey, "Horn Antennas with Uniform Power Patterns Around Their Axes," IEEE Trans., *AP-14* (1966), pp. 656–658.
2. H. C. Minnet and B. M. Thomas, "A Method of Synthesizing Radiation Patterns with Axial Symmetry," IEEE Trans., *AP-14* (1966), pp. 654–656.
3. P. J. B. Clarricoats and P. K. Saha, "Propagation and Radiation Behaviour of Corrugated Feeds, Part I—Corrugated Waveguide Feed," Proc. IEE, *118,* No. 9 (September 1971), pp. 1167–1176.
4. C. Dragone, "Characteristics of a Broadband Microwave Corrugated Feed: A Comparison Between Theory and Experiment," B.S.T.J., *56,* No. 6 (July–August 1977), pp. 869–888.
5. T. S. Chu and R. H. Turrin, "Depolarization Properties of Off-set Reflector Resonators," IEEE Trans., *AP-21* (May 1973), pp. 334–345.
6. M. J. Gans, "Cross-Polarization in Reflector-Type Beam Waveguides and Antennas," B.S.T.J., *55,* No. 3 (March 1976), pp. 289–316.
7. Hirokaau Tanaka and Motoc Mizusawa, "Elimination of Cross-Polarization in Offset Dual-Reflector Antennas," Electronics and Communications in Japan, *58-B,* No. 12 (1975).
8. R. Graham, "The Polarization Characteristics of Offset Cassegrain Aerials," IEEE International Conference on Radar—Present and Future, No. 105 (October 1973). Also United States Patent No. 3,792,480.
9. Y. Mizugutch, M. Akagawa, and H. Yokoi, "Offset Dual Reflector Antenna," Digest of 1976 AP-S International Symposium on Antennas and Propagation, October 1976.
10. K. N. Coyne, unpublished work.
11. A. B. Crawford, D. C. Hogg, and L. E. Hunt, "A Horn-Reflector Antenna for Space Communication," B.S.T.J., *40,* No. 4 (July–August 1961), pp. 1005–1116.
12. M. Mizasawa and T. Kitsuregawa, "A Beam-Waveguide Feed Having A Symmetric Beam for Cassegrain Antennas," IEEE Trans. Ant. Propag., *AP-21,* No. 6 (November 1973), pp. 884–886.
13. M. Mizusawa and T. Katoigi, "A Property of the Series of Mirrors on Quadratic Surface of Revolution," Trans. IEEE Japan, *53-B* (November 1970), pp. 707–708.
14. C. Dragone, "New Grids for Improved Polarization Diplexing of Microwaves in Reflector Antennas," IEEE Trans. Ant. Propag., *AP-26,* No. 3 (May 1978), pp. 459–463.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 9, September 1978
Printed in U.S.A.

Radiation Patterns from Parallel, Optical Waveguide Directional Couplers—Parameter Measurements

By V. RAMASWAMY and R. D. STANDLEY

(Manuscript received November 22, 1976)

A new method for measuring the parameters of optical, parallel, waveguide directional couplers is presented. Basically, we observe the changes in radiation pattern obtained by placing a high, refractive, index coupling prism on the coupled guides as a function of position along the coupler. For a coupler, in a Z cut, Ti-diffused $LiNbO_3$ substrate with 3-μm guides and 3-μm separation, the transfer length is about 1.8 mm at 7266 Å.

I. INTRODUCTION

Parallel coupled waveguides are the basic building block for a number of integrated optical devices; these include switches,[1–6] modulators, and channel dropping or adding filters.[7] The techniques used to measure coupling parameters are often visual in nature. The simplest approach is to observe the energy exchanges between the parallel guides from the surface scattering of these guides viewed through a microscope. However, this is not always feasible; e.g., operation at longer wavelengths away from the visible, with low-loss surface scattering guides, and in cases where the energy at the surface is rather low, as it happens with Ti diffused guides in $LiNbO_3$. In such cases, the technique developed by Ostrowsky et al.[8] is quite useful. They observed the fluorescence from RhB-doped polyurethane film over the strip guides pumped by an argon laser.

In this paper, we present a method found useful in measuring the parameters of such couplers. Basically, the method consists of observation of the interaction length dependence of the coupling via radiation pattern measurements;[9] the radiation patterns are obtained by moving an output coupling prism along the coupled waveguide region.

II. THEORY

2.1 Synchronous couplers

Figure 1 depicts two coupled parallel waveguides where a is the guide width, c is the guide spacing, and L is the length over which the guides are coupled; i.e., the interaction length. We consider the ideal case which assumes that the guides are identical in width and thickness so that perfect synchronism of the unperturbed propagation constants exists; for this case, the normalized field amplitudes in the two guides as a function of length z can be shown to be[10]

$$\begin{aligned} R &= \cos \kappa z \\ S &= j \sin \kappa z, \end{aligned} \tag{1}$$

where R is the field amplitude in the initially excited guide, S is that of the auxiliary guide, and κ is the coupling strength per unit length. We are interested in determining the coupling strength κ per unit length for a coupler of known physical parameters; knowledge of κ permits the selection of L for a coupler of desired overall coupling strength. If, at some point z along the parallel coupled region, we place a prism whose refractive index is higher than that of the waveguides, then power will be radiated from the two waveguides. Thus in the far field we observe a radiation pattern due to the interference of the fields from the coupled waveguides over the coupling length of the prism coupler. If we keep the prism coupling length small compared to $1/\kappa$, say, less than a millimeter, then the far-field radiation pattern would truly be representative of the pattern from two slits separated by a distance of d having relative amplitudes given by eq. (1).

If we assume constant transverse field amplitudes, as seen from Fig. 2, the expression for the radiation pattern is

$$|E|^2 = \frac{\sin^2 u}{u^2}\left(1 + \sin 2Z \sin 2u \frac{d}{a}\right), \tag{2}$$

where

$$Z = \kappa z$$

$$u = \frac{\pi a}{\lambda} \sin \theta$$

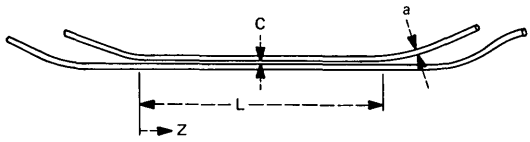
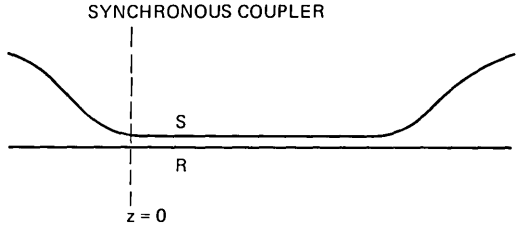
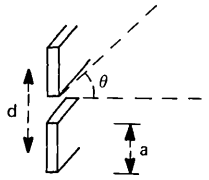


Fig. 1—Parallel waveguide directional coupler where a is guide width, c is the guide spacing, and L is the interaction length.





$$E = \left(S + Re\ \frac{j2\pi d}{\lambda}\sin\theta\right)\frac{\sin u}{u}$$

$$E^2 = \frac{\sin^2 u}{u^2}\left(1 + \sin 2\kappa z \sin\frac{2ud}{a}\right)$$

$$u = \frac{\pi a}{\lambda}\sin\theta$$

Fig. 2—Radiation field amplitudes under far-field conditions due to a pair of sources of width a, separated by a distance d. θ is measured in the plane perpendicular to the plane containing the waveguides.

$$d = a + c$$

θ is radiation angle.

Figure 3 shows computed plots of $|E|^2$ as a function of u for the case $d/a = 2$ with Z as the parameter. Except when all of the energy is in one guide, e.g., at $Z = o$, the radiation pattern is asymmetrical about $\theta = o$. This is true even for the case when $Z = \pi/4$, when the field amplitudes in both guides are equal, and differ by a phase shift of 90 degrees. When Z is increased from $\pi/4$ to $\pi/2$ in specific increments, the patterns remain the same as Z is varied from $\pi/4$ to 0, for the same shape, i.e., for example, identical patterns are observed for the cases when $Z = \pi/16$ and $7\pi/16$, $\pi/8$ and $3\pi/8$, $3\pi/16$ and $5\pi/16$, etc. At $Z = Zo = \pi/2$, complete energy transfer occurs. When Z is varied from $\pi/2$ to $3\pi/4$ and back to π, the graphs shown in Fig. 3 can be used with change in sign of abscissa. The whole series of patterns repeat themselves in this manner with increasing Z.

2.2 Asynchronous couplers

If the waveguides differ in width, thickness, or refractive index, their propagation constants will differ. This could occur as a result of errors in the fabrication process. For such asynchronous couplers, complete power transfer from one guide to the other is not possible. If we define

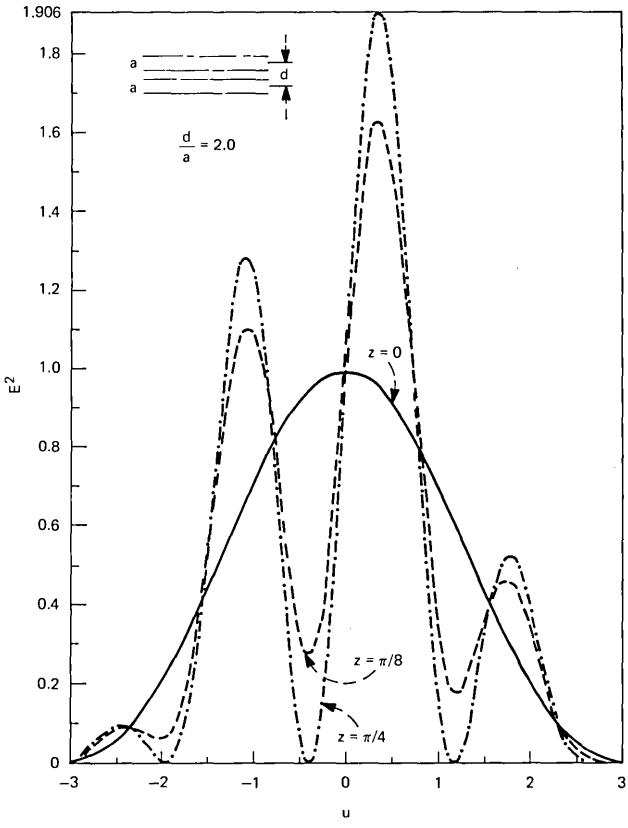


Fig. 3—Computer plots of energy distribution as a function of normalized radiation angle u for the case $d/a = 2$, with $Z = \kappa z$ as the parameter.

the difference in unperturbed propagation constants in the two guides as $\Delta\beta$, then the normalized field amplitudes[6] as a function of z become

$$R' = \cos\alpha - j\frac{\gamma}{\sqrt{\gamma^2+1}}\sin\alpha$$
$$S' = j\frac{\sin\alpha}{\sqrt{\gamma^2+1}}, \tag{3}$$

where

$$\gamma = \Delta\beta/2\kappa$$
$$\alpha = \sqrt{\gamma^2+1}\,\kappa z.$$

Here, again, R' is the field amplitude in the initially excited guide and S' is that of the auxiliary guide. With these field amplitudes, the radia-

tion pattern is given by

$$|E'|^2 = \frac{\sin^2 u}{u^2}\left(1 + \frac{\sin 2(\gamma^2+1)^{1/2}Z}{(\gamma^2+1)^{1/2}}\sin\left(2u\frac{d}{a}\right)\right.$$
$$\left. - \frac{\gamma}{(\gamma^2+1)}[1 - \cos 2(\gamma^2+1)^{1/2}Z]\cos\left(2u\frac{d}{a}\right)\right). \quad (4)$$

The power in the coupled guide is obtained by squaring eq. (3) and is given by

$$|S'|^2 = \frac{\sin^2[(\gamma^2+1)^{1/2}Z]}{(\gamma^2+1)}$$

and (5)

$$|R'|^2 = 1 - |S'|^2.$$

We find the maximum value for the coupled power to be $(\gamma^2+1)^{-1}$ at $Z = (m\pi/2)(\gamma^2+1)^{-1/2}$. Plots of $|E'|$ show the expected result that the information content in the radiation patterns decreases rapidly with increasing asynchronism. However, useful information is obtained by recognizing the transfer period as indicated by all the power being present in the input guide.

III. COUPLER FABRICATION AND MEASUREMENT TECHNIQUE

The procedures used in the fabrication of the experimental couplers are described. Z-cut lithium niobate substrates were coated with polymethyl-methacrylate (PMMA) electron resist approximately 0.5 micron in thickness. A thin layer of aluminum (100 Å) is evaporated onto the

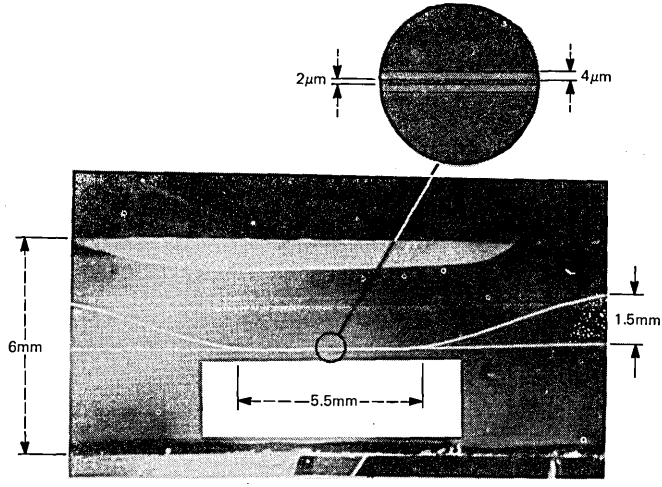


Fig. 4—Guide tracks defined in PMMA after electron beam exposure and development.

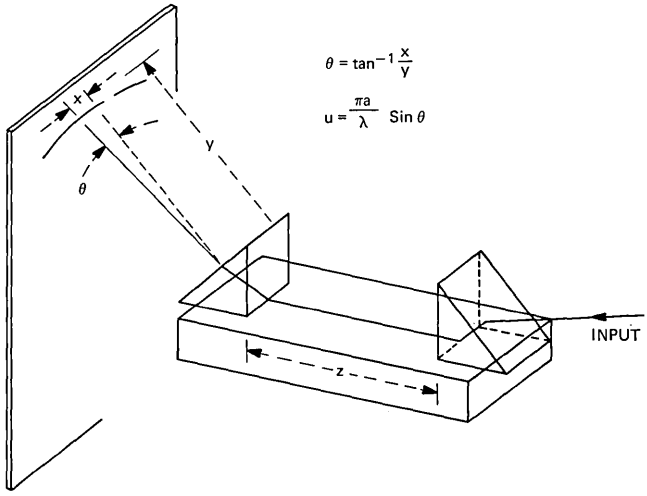


Fig. 5—Illustration of the setup to measure the coupler radiation pattern.

PMMA to eliminate charging problems. The coated substrate is then mounted onto a scanning electron microscope (SEM) stub using a conducting silver paste. Using the appropriate scan generator, the first guide of the coupler is exposed. The scan generator output amplitude is then attenuated and the writing beam moved by electronic adjustment of the fine shift coil current; the auxiliary guide is then exposed. For exposure, a specimen current of 10^{-9} A is typically used with an exposure time of about 25 s to obtain 3-μm wide guides 15 mm in length. The sample is then removed from the SEM. A brief soak in dilute NaOH removes the aluminum layer. The PMMA is then developed for about 30 s in a 3-to-1

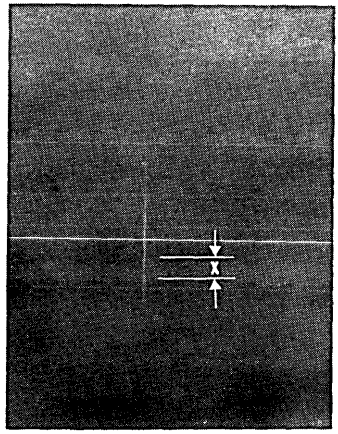


Fig. 6—A typical radiation pattern—in this case, the energy is very close to the position where all the energy is one of the guides.

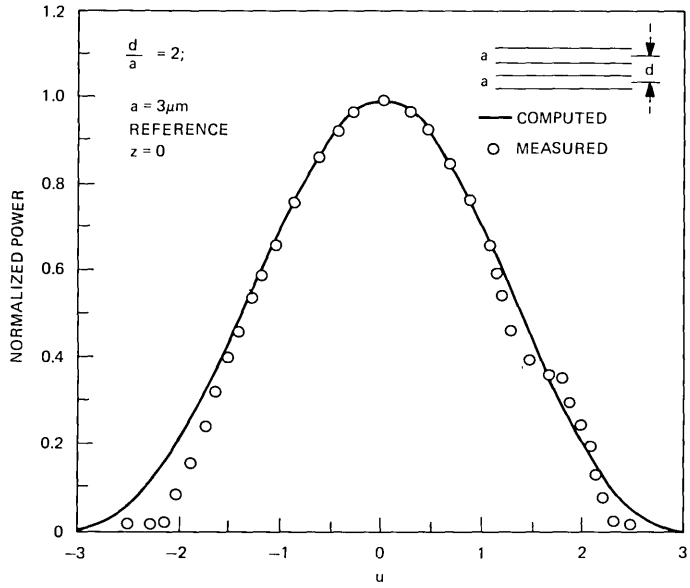


Fig. 7—Radiated power output as a function of normalized radiation angle u where all the energy is in one of the guides, $z = 0$. $d/a = 2$ for this coupler.

mixture of isopropyl alcohol and methylethyl ketone. The guide tracks are now defined in the PMMA (Fig. 4). The sample is blown dry with dry nitrogen and mounted in a sputtering system for deposition of a Ti layer usually about 300 Å thick. The PMMA and excess Ti are next removed by soaking the sample in acetone. At this point, we have a sample with Ti where we want the waveguides. The sample is next placed in an oven and brought to 1000°C in an argon ambient. Following the 1000°C soaking for about three hours, the furnace is turned off and the ambient changed to oxygen. The resulting guides exhibit single TE mode operation.

The experimental set-up used to measure the coupler radiation pattern is shown in Fig. 5. The lasers employed were He-Ne operating at 6328°A and a nile-blue dye laser covering the wavelength 6900°A to 7500°A. The latter source was pumped by the 6471°A line of a krypton laser. The prisms were made of rutile. The input prism was quite flat, allowing strong coupling, whereas the base of the output prism had a curvature in it to ensure the coupling region to be much less than that of a millimeter. Although the amount of energy coupled out is rather small, the resulting radiation pattern is primarily due to the energy of the guide at the output prism location and does not include the effects of long coupling lengths. As the output prism was moved along the guides, the radiation pattern was scanned using an iris. Figure 6 is a photograph of a typical nearly synchronous coupler radiation pattern. The pattern in Fig. 7 resulted from a coupler operating at 7266°A con-

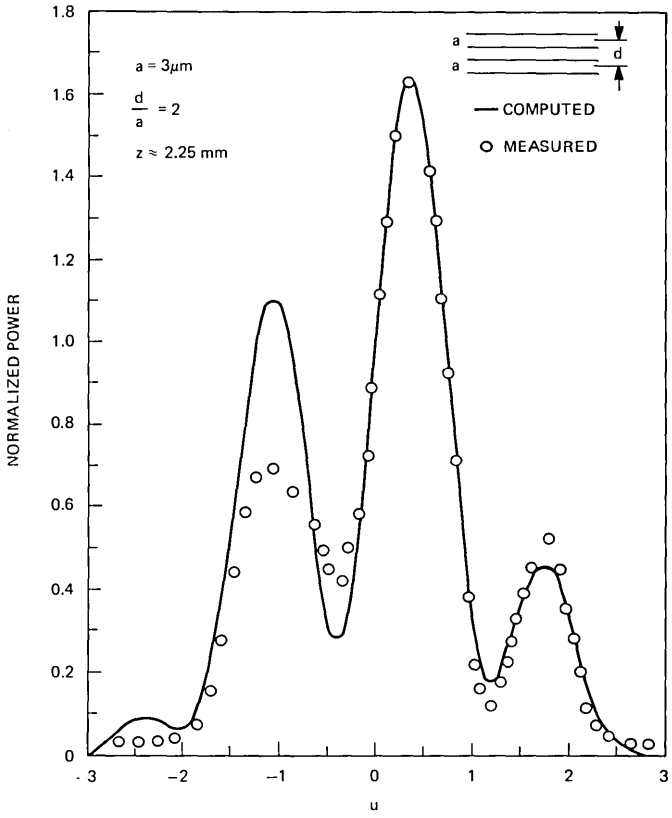


Fig. 8—Radiation pattern of the same coupler shown in Fig. 7, but at a different $z \simeq 5\pi/8$, past the location of complete energy transfer.

sisting of 3-μm guides. The measurement was made at a position $Z = 0$ along the coupler. Where all power was essentially in one of the guides, measured distribution agrees well with the theory. By moving the output prism to a place where all the energy is in the other guide, the transfer length can be measured. However, if the prism is not placed exactly at this location, one can infer this information by noting the nature of the asymmetry and measuring the radiation pattern. For example, Fig. 8 shows the output radiation pattern for the same coupler, but at a different longitudinal position $z = 2.25$ mm. In this case, the power in the two guides is nearly equal, resulting in sidelobe development in the observed radiation pattern. From Fig. 3 for $d/a = 2.0$, the separation Δu between minima is 1.6. At 7266°A, for $a = 3$ μm, this translates into a separation $\Delta x = 2.24$ cm between minima at a distance $y = 18$ cm from the output prism coupling position. This compares favorably with the measured value of 2.25 cm. By a series of observations on this coupler, we can infer an interaction length for full power transfer $L_o = \pi/2\kappa$; the

best fit for curve in Fig. 8 occurs at $Z = \pi z/2L_o = 5\pi/8$, from which the transfer length L_o is inferred to be 1.8 mm for this coupler. The separation of the minima agrees very well, although the peaks do not. Considering that we analyze uniform distribution of energy in the waveguides, the agreement is rather good.

IV. CONCLUSION

We have described a method for measuring the coupling strength of synchronous optical waveguide directional couplers by observing the length dependence of the radiated signal. As indicated earlier, the technique is useful, with care in implementation, as a laboratory tool.

V. ACKNOWLEDGMENTS

The authors are grateful to M. D. Divino and F. A. Braun for their assistance, and to Mrs. D. Vitello for programming the computer plots presented in this paper.

REFERENCES

1. H. F. Taylor, "Optical Switching and Modulation in Parallel Dielectric Waveguides," J. Appl. Phys., *44*, no. 7 (1973), p. 3257.
2. F. Zernike, "Integrated Optical Switch," WA5 Topical meeting on Integrated Optics, New Orleans, 1974.
3. M. Papachon et al, "Electrically Switched Optical Directional Couplers: Cobra," Appl.Phys. Lett., *27* (Sept. 1975), p. 289.
4. J. C. Campbell et al, "GaAs Electrooptic Directional Coupler Switch," Appl. Phys. Lett., *27* (August 1975), p. 202.
5. H. Kogelnik and R. V. Schmidt, "Switched Directional Couplers with Alternating $\Delta\beta$," IEEE J. Quantum Electronics (July 1975).
6. V. Ramaswamy and R. D. Standley, "A Phased, Optical, Coupler Pair Switch," B.S.T.J., *55*, No. 6 (July-August 1976), p. 767.
7. V. Ramaswamy and R. D. Standley, patent pending.
8. D. B. Ostrowsky et al., Appl. Opt., *13* (March 1974), p. 636.
9. R. D. Standley and V. Ramaswamy, "A New Method for Measuring Parallel Waveguide Directional Coupler Parameters," MD3 Topical Meeting on Integrated Optics, Salt Lake City, January 1976.
10. S. E. Miller, "Coupled Wave Theory and Waveguide Applications," B.S.T.J., *33*, No. 3 (May 1954), p. 661.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 7, September 1978
Printed in U.S.A.

Speech Signal Power in the Switched Message Network

By W. C. AHERN, F. P. DUFFY, and J. A. MAHER

(Manuscript received December 15, 1977)

Speech signal power at the main distributing frame in class 5 switching offices is characterized in terms of equivalent peak level (EPL) and average conversational signal power measures. The results indicate that there is little dependence of speech signal power on call destination or originating class of service. Small differences between various subpopulations are explained for the most part by loop characteristics. The switched telecommunications network is essentially transparent to customers in the sense that talker signal power has not been found to be sensitive to factors which affect the transmission path between class 5 central offices.

Present-day speech volumes for toll calls, which average −21.6 VU (volume units), are substantially lower than those found in a survey conducted in 1960,[1] which averaged −16.3 VU, and the ranges of volumes within all call destination categories are substantially smaller than the 1960 ranges. Several substantial changes have been introduced into the telephone plant since 1960 which tend to increase the uniformity of service in the network from the viewpoint of speech volumes. These include a decrease in the proportion of toll grade battery, loss plan improvements, replacement of the 300-type telephone set with the 500-type set, and an increase in direct trunking between class 5 offices.

I. INTRODUCTION

The characterization of speech signal power on Bell System message circuits is an essential step in the determination of signal power loading and crosstalk objectives. Knowledge of speech signal characteristics is also important to designers of a wide variety of telecommunications equipment.

Speech levels at the class 5 office were last characterized in the 1960 Speech Volume Survey[1] in terms of volume units (VU). In the years since

the last survey, there have been substantial changes in the Bell System network. For example, the proportion of toll grade battery has been substantially reduced, the 300-type telephone set has been almost completely phased out, direct distance dialing is now virtually universal, and new loop and trunk design methods have been introduced. Also, in the intervening years, research in speech signal measurement has led to a new measure of speech level known as the equivalent peak level (EPL).[2] This, together with advanced digital data acquisition technology, has facilitated the measurement of speech signal power with greater precision than was possible in 1960.

This paper presents the results of a speech signal power survey made in 1975–1976. The measurements were made at 36 class-5-office main distributing frames (MDFs), which constitute a statistical sample of acceptable precision from all the MDFs within the Bell System. The class 5 (local or end) office MDF was selected as the measurement interface because it has access to all customer loops and all classes of local and toll traffic; dialed address information is readily available; only the customer's loop and station equipment is interposed between the customer and the point of measurement; and the customer's loop current may be measured. A three-stage statistical sampling scheme was employed, which resulted in measurements of near-end and far-end talker power on more than 10,000 calls originating from approximately 2500 loops. Average conversational signal power (averaged over the entire observation interval) and EPL were the measures used for talker signal characterization. Loop dc current, class of service, switch type, and call destination were also recorded.

Survey results are presented in Section II, the methodology is presented in Section III, and comparisons of the present survey results with prior survey results are given in Section IV.

II. SURVEY RESULTS

Table I summarizes the findings of this survey. The results indicate that there is little dependence of speech signal power on call destination or originating class of service. In the sections that follow, it is shown that the small differences between various subpopulations are explained for

Table I—Summary of speech signal powers

Subclass	Near-End Mean Equivalent Peak Level (dBm)
Residence	−11.0
Business	−10.4
Local	−10.8
Toll	−10.1
Combined	−10.7

the most part by loop characteristics, and there is little if any variation in speech signal power that may be attributable to psychological factors such as call distance, perception of received volume, etc. The indication from the data is that the switched telecommunications network is essentially transparent to customers in the sense that talker signal power has not been found to be sensitive to call distance, local or toll call classification, or other factors that affect the transmission path from class 5 to class 5 central office.

2.1 General

In this survey, speech signal power measurements were made on customer loops at class 5 switching office main distributing frames (MDFs) during actual telephone conversations. The parties originating calls on sampled loops are referred to as the "near-end" speakers in the following discussion; the called parties are referred to as the "far-end" speakers. The far-end speakers were more distant than the near-end speakers from the MDFs at which the measurements were made, except for some intrabuilding calls.

The survey results characterize near- and far-end speech signal powers in terms of the equivalent peak level (EPL) and average conversational signal power measures, which are discussed in Section 3.3.3. The differences are also characterized between near- and far-end signal powers and between the EPL and average power measures. In addition, the influences of loop current, originating class of subscriber service, call destination, call distance, originating switch, and demographic features upon speech signal powers are investigated.

2.2 Speech signal powers at main distributing frames

The distributions of speech signal power at main distributing frames can be approximated by normal distributions. Histograms and cumulative distribution functions (CDFs) are given for the EPL and average power measures of speech signal power for the near- and far-end speakers in Figs. 1 through 4. The "bell" shapes of the histograms and the straight line shapes of the CDFs, which are plotted on normal probability grids, attest to the normality of these distributions. Because of this, the distributions are completely defined by the means and standard deviations listed in the first four lines of Table II.

While the near- and far-end signals encounter similar populations of station set and subscriber loop losses, the far-end signals also encounter end-office-to-end-office transmission losses. As a result of these additional losses, which will be referred to as the "apparent network loss" during the remainder of this paper, the average far-end signal power is generally lower than the average near-end signal power. The apparent network loss is a function of call destination, i.e., the greater the call

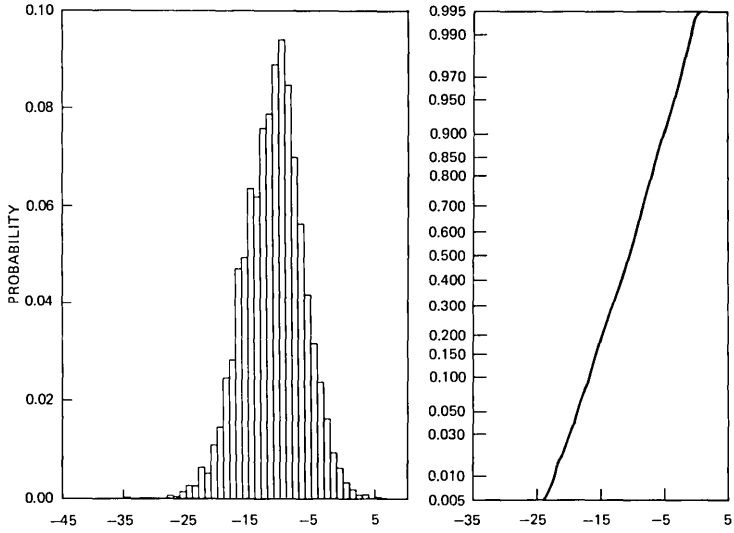


Fig. 1—Near-end equivalent peak level (dBm) distribution.

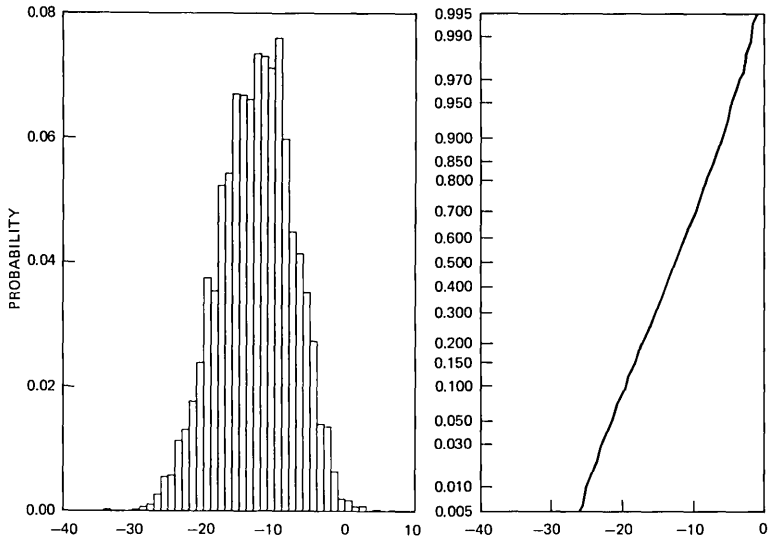


Fig. 2—Far-end equivalent peak level (dBm) distribution.

distance between end offices the more the signals are attenuated. This source of variation explains the greater variability among the far-end signal powers. These near-end, far-end differences exist for both EPL and average power; however, a comparison of the near- and far-end EPL results gives a difference of 2.1 dB, while a similar comparison for the average power measures gives a difference of 2.9 dB. In the following

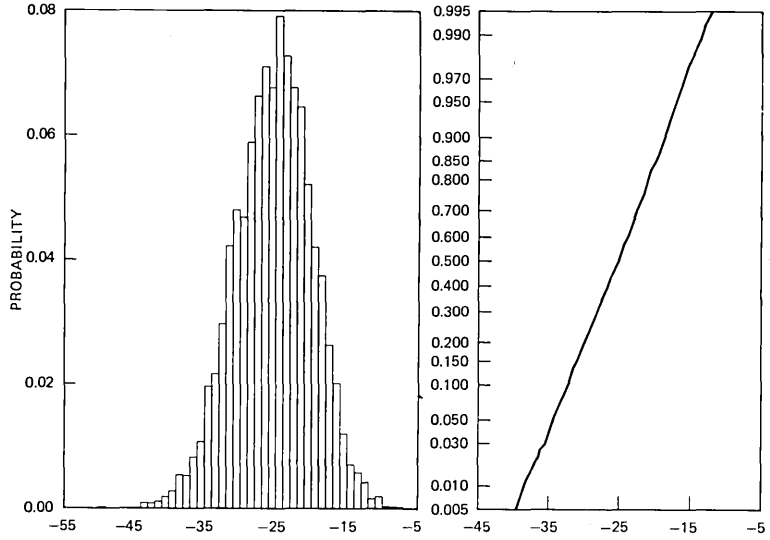


Fig. 3—Near-end average signal power (dBm) distribution.

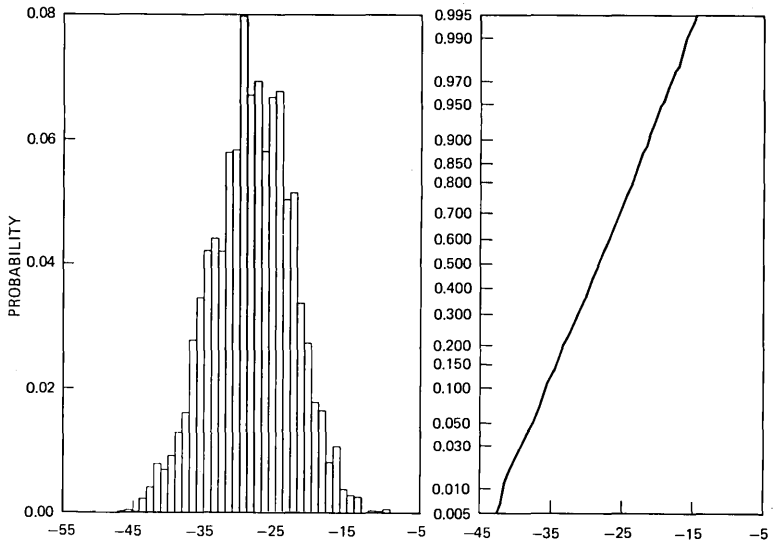


Fig. 4—Far-end average signal power (dBm) distribution.

paragraph, this apparent anomaly is shown to be caused by a difference in the speech activity of near- and far-end speakers.

The EPL, which is derived from the speech samples exceeding a threshold, is a measure of the speaker's peak signal power, and therefore is unaffected by silent periods in the conversation. The average signal power for conversational speech, however, includes intervals of speech

and silence alike. Therefore, the average power measure is lower than the corresponding EPL. This is illustrated by the results in Table II, which show that the average difference between EPL and average power is 14.6 dB for the near-end measures and 15.6 dB for the far-end measures. Such differences represent activity factors in the sense that they are logarithmically related to the amount of silence during a conversation.[3] They indicate that calling parties (near-end) tend to speak more than called parties (far-end) during telephone conversations. Due to these different speech activity characteristics, the apparent network loss result based upon average power is overestimated by about 1 dB. This finding explains the apparent anomaly noted above, and suggests that EPL is more appropriate than average power for estimating apparent network loss.

Comparisons of near-end EPL and average power with the far-end measurements are provided in the scatter diagrams in Figs. 5 and 6. The correlation coefficients are 0.31 and 0.57 for the EPL and average power comparisons, respectively. While the relationships are statistically significant, the modest positive correlations indicate that the signal power of one speaker is not strongly influenced by the signal power of the other.

Average signal power is strongly related to EPL. The results of the linear regressions of the near- and far-end EPLs on the corresponding average powers are given in Figs. 7 and 8, respectively. The near-end regression shows that average power = −14.27 + 1.04 EPL, and the far-end regression shows that average power = −15.40 + EPL. The values of R^2, the square of the correlation, on the figures indicate that approximately 85 percent of the variation in average signal power is accounted for by the regression fits.

Signal power at the MDF is dependent upon loop loss and the telephone set electroacoustic efficiency. While these parameters were not measured, the near-end loop current, which was measured, has been found to relate to the overall loop and telephone set loss.[4] The histogram

Table II—Systemwide speech signal power results

Transmission Characteristic	Mean	90% C.I.	Std. Dev.	Sample
Near-end EPL (dBm)	−10.7	±0.5	4.6	10251
Far-end EPL (dBm)	−12.7	±0.4	5.2	8976
Near-end average power (dBm)	−25.3	±0.5	5.3	10251
Far-end average power (dBm)	−28.3	±0.4	5.6	8976
Near minus far-end EPL (dB)	2.1	±0.4	5.9	8478
Near minus far-end average power (dB)	2.9	±0.4	6.7	8478
Near-end EPL minus average power (dB)	14.6	±0.1	2.1	10251
Far-end EPL minus average power (dB)	15.6	±0.1	2.1	8976
Near-end loop current (mA)	42.2	±1.9	12.8	10749

90% C.I. = 90-percent confidence interval for the mean estimate.
Std. Dev. = Standard deviation of the signal power or loop current population.
Sample = Total sample size in calls used to calculate estimates.

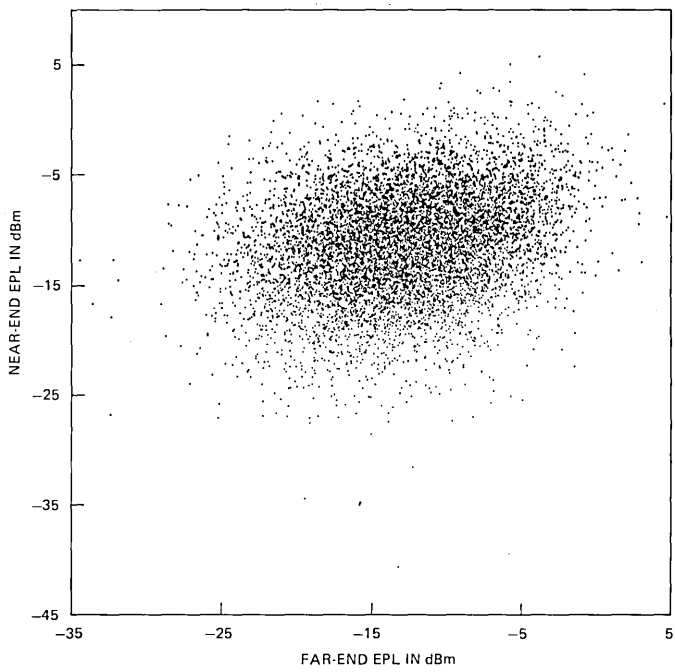


Fig. 5—Comparison of near- and far-end equivalent peak level.

and CDF for loop current are given in Fig. 9. The distribution is positively skewed, which means that it deviates from normality due to some large values of loop current associated with short loops. The distribution also deviates from normality at the lower tail because of a truncation of loop currents below 20 mA due to engineering limitations for signaling and transmission systems. Table II shows that the average loop current is 42.2 mA and the standard deviation is 12.8 mA.

Near-end EPL and average power are plotted as a function of loop current in Figs. 10 and 11, respectively. The scatter diagrams indicate that EPL and average signal power increase as loop current increases. Loop and telephone set characteristics suggest that a nonlinear relationship exists between loop current and the total loop and telephone set loss.[4] Nonlinear regression confirms this; however, the improvement in fit over the linear model, while statistically significant, is not of practical interest. The linear regressions of EPL and average power on loop current indicate that signal power increases about 0.13 dB per 1.0 mA increase in loop current. However, signal power varies substantially about the regression lines, indicating that loop current alone is not a good predictor of signal power. Visually, the variance appears to depend upon loop current; however, an analysis within loop current categories indicates that the variance is constant.

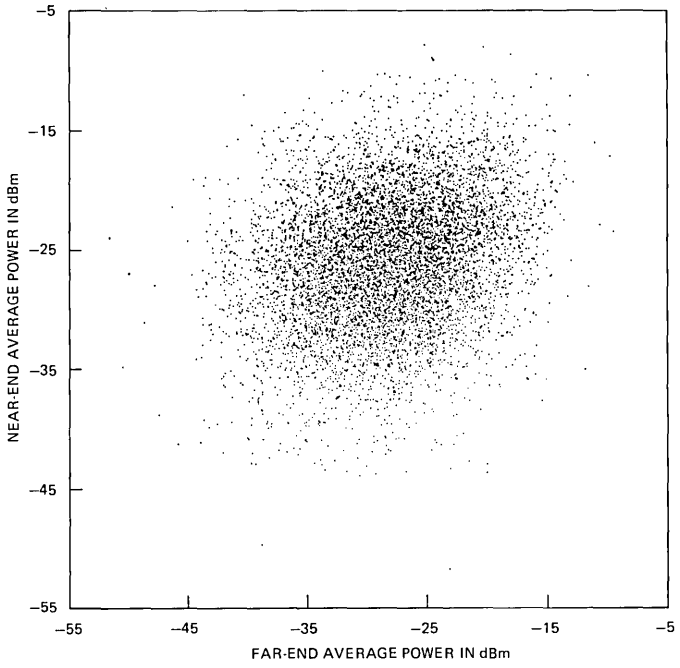


Fig. 6—Comparison of near- and far-end average power.

A more vivid illustration of the relationship between signal power and loop current is given in Fig. 12 by plotting the average EPL for each of the 36 MDFs in the sample as a function of the average loop current per MDF. The scatter shows a positive correlation, and the correlation coefficient is 0.82. A linear regression indicates that average EPL = −19.06 + 0.20 average loop current, and that the regression fit accounts for 67 percent of the variability in average EPL among MDFs.

2.3 Signal power and class of service

Class of service identifies the subscriber as a business or residential customer and identifies the station terminals as Bell or customer-provided equipment (CPE). The analyses discussed in this section deal with these service perspectives on the basis of originating class of service. The terminating customer class of service was not determined for the calls in this survey.

2.3.1 Business versus residential

The survey results for business- and residential-originated calls are summarized in Table III. Comparisons of the near-end EPL and average power results indicate that business-associated signal powers tend to be slightly higher than residential-associated signal powers, and that

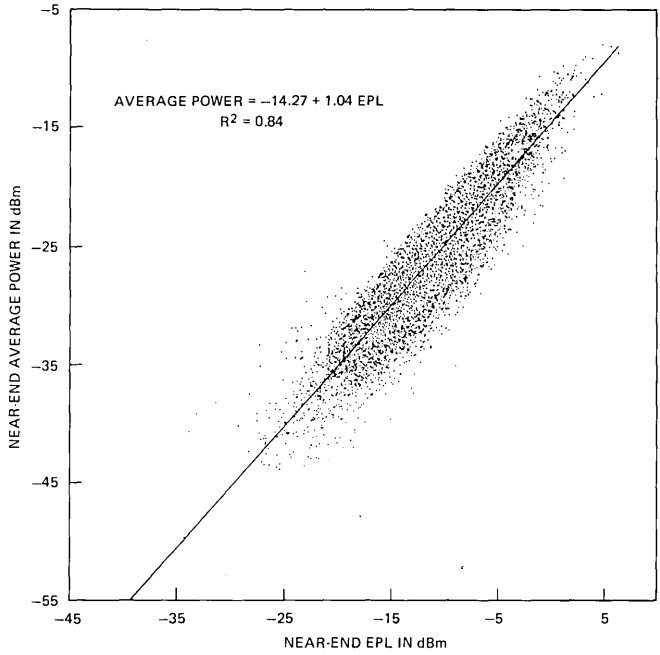


Fig. 7—Linear regression of near-end average power on equivalent peak level.

the variability among signal powers is about the same in both service categories. The 90-percent confidence intervals for the business and residential averages overlap, indicating that the differences are not statistically significant. Business loop currents are significantly higher and more variable than residential loop currents. The 5.3-mA difference in average loop current combined with the finding in Section 2.2, which indicates that EPL increases 0.13 dB per 1.0 mA increase in loop current, suggests that the business average EPL should be about 0.7 dB higher than the residential average. This difference agrees with the residence-business difference found for the near-end talker.

The far-end signal power results derived from the analysis by originating class of service are almost identical in the business and residential classifications. Since the originating parties in either category place calls to business and residential stations alike, the far-end speakers in each originating class of service category represent a mixture of business and residential customers. The far-end class of service mixture within each originating class of service category is sufficiently close to the overall traffic composition that the far-end results in each category are essentially the same as the far-end results for all telephone traffic listed in Table II. It is interesting to note that, although the average calling dis-

Table III—Originating class of service speech signal power results

Transmission Characteristic	Business				Residential			
	Mean	90% C.I.	Std. Dev.	Sample	Mean	90% C.I.	Std. Dev.	Sample
Near-end EPL (dBm)	−10.4	±0.7	4.6	6072	−11.0	±0.4	4.7	4179
Far-end EPL (dBm)	−12.8	±0.5	5.2	5228	−12.7	±0.4	5.2	3748
Near-end average power (dBm)	−25.0	±0.8	5.2	6072	−25.7	±0.4	5.4	4179
Far-end average power (dBm)	−28.4	±0.5	5.5	5228	−28.2	±0.4	5.6	3748
Near minus far-end EPL (dB)	2.5	±0.7	6.0	4916	1.7	±0.3	5.8	3562
Near minus far-end average power (dB)	3.4	±0.7	6.7	4916	2.5	±0.3	6.6	3562
Near-end EPL minus average power (dB)	14.6	±0.2	2.1	6072	14.7	±0.1	2.1	4179
Far-end EPL minus average power (dB)	15.7	±0.1	2.1	5228	15.5	±0.1	2.1	3748
Near-end loop current (mA)	45.0	±2.9	13.9	6384	39.7	±1.2	11.1	4365

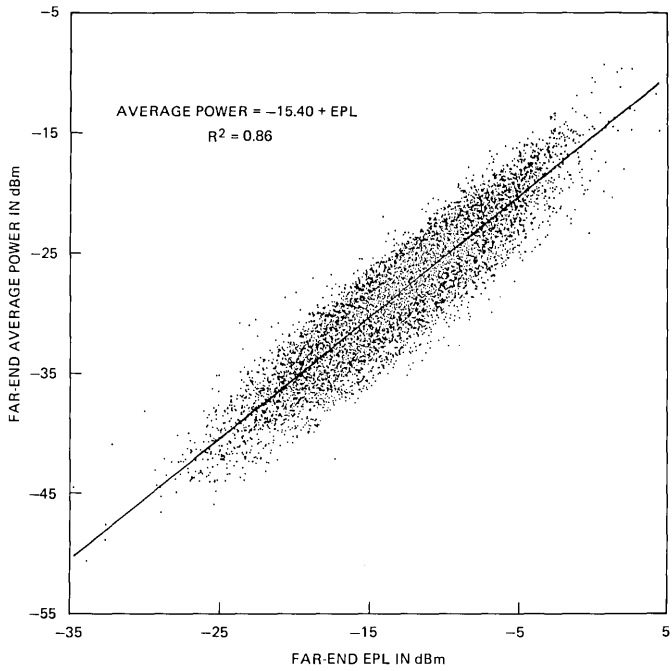


Fig. 8—Linear regression of far-end average power on equivalent peak level.

tance for the business-originated calls (50 ± 12 miles) is over 3.5 times the average for the residential calls (14 ± 4 miles), there is no noticeable call distance impact upon far-end talker received signal power. This does not imply that call distance has no influence upon network loss; it does imply that most of the data represent local calls or very short toll calls, and thus any potential call distance influence is not apparent.

Speaker speech activity during a telephone conversation is not affected by originating class of service. The EPL-average power differences have similar distributions for business- and residential-originated conversations.

The signal power distributions are all close to normal for business and residential calls. Therefore, the EPL and average power results listed in Table III completely define the signal power distributions for all practical applications. The business and residential loop current distributions differ significantly and are presented in Figs. 13 and 14, respectively. The business loop current distribution is comparable to the 1964 General Loop Survey[5] computed loop current distribution. The residential distribution has a greater proportion of lower current loops than the 1964 Survey result.

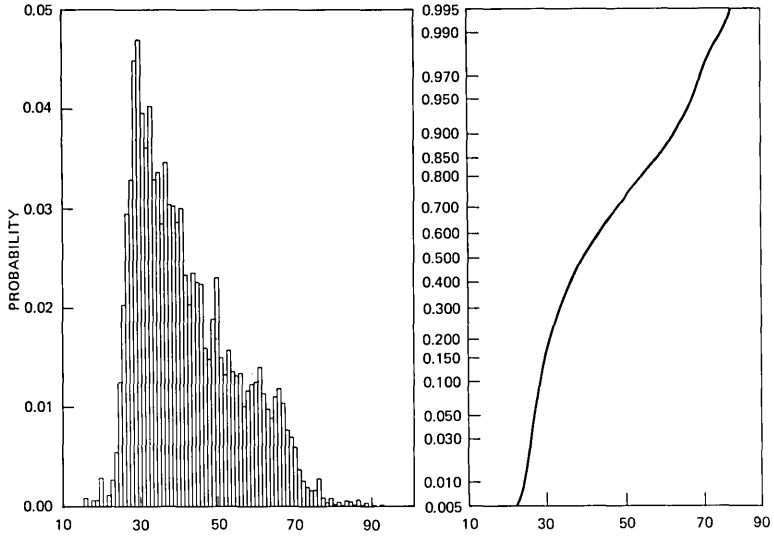


Fig. 9—Near-end loop current (mA) distribution.

2.3.2 Bell versus customer-provided equipment

Business calls are further classified on the basis of terminal equipment ownership in this section. One category contains those business calls which originated from subscriber lines with terminal equipment leased from the Bell System, and the second category contains those calls which originated from subscriber lines with customer-provided equipment (CPE). The results of this analysis are tabulated in Table IV. The near-end estimates show that the Bell signal powers on the average are more than 2 dB higher than the CPE signal powers, and that they are also somewhat less variable. The reason for this difference is suggested by examining the relationship between loop current and EPL for Bell and CPE loops, respectively. The correlation coefficients are 0.39 and 0.16, respectively, indicating that speech signal power on CPE loops is less strongly influenced by loop current than in the case of Bell loops. The reason for this is that the CPE station equipment battery is provided by a local power supply and not over the metallic loop facility. Thus, the electroacoustic efficiency of CPE station equipment is unrelated to the loop current observed in the central office, and the lower mean and higher variance in signal power may be attributable to the various local battery supplies and electroacoustic efficiencies of CPE terminals.

Comparisons of the far-end signal power estimates indicate that those far-end signals associated with CPE-originated calls have slightly lower signal powers than those associated with Bell-originated calls. The absence of detailed information about the far-end customers prevents further analyses to determine the cause of this difference.

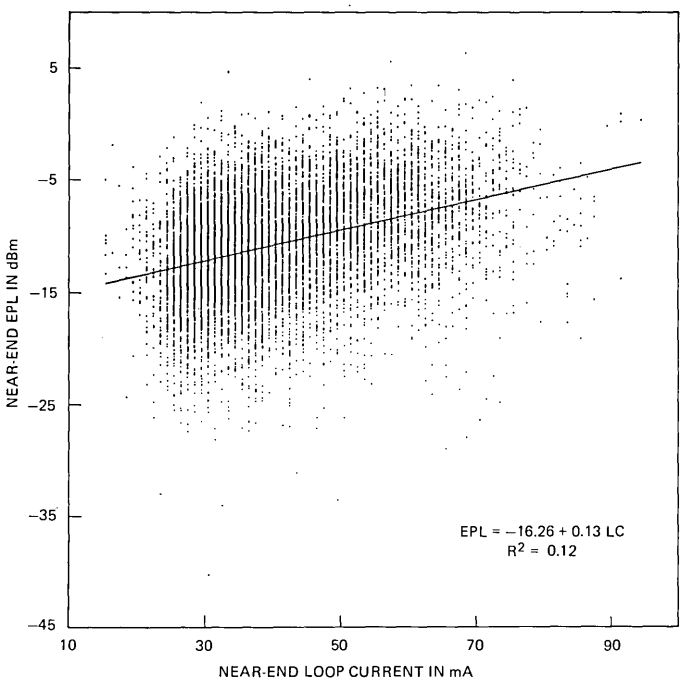


Fig. 10—Linear regression of equivalent peak level on loop current.

The signal power distributions again are all close to normal for both Bell- and CPE-originated business calls. The loop current distributions for both categories are comparable to the distributions given in the previous section for business calls in general.

2.4 Signal power and call destination

Four categories of call destination are considered in the following discussion; (*i*) intrabuilding local calls, (*ii*) interbuilding local calls, (*iii*) Home Numbering Plan Area (HNPA) toll calls, and (*iv*) Foreign Numbering Plan Area (FNPA) toll calls. The first two of these categories characterize local calls, and the last two characterize toll calls.

The trend lines in Fig. 15 summarize the relationships between signal power and call destination and between loop current and call destination. The near-end EPL and average power appear to increase slightly as the call destination becomes more remote from the originating office, with the exception of a slight drop in signal power for interbuilding local calls. The 90-percent confidence intervals for the four EPL estimates and for the four average power estimates overlap, which indicates that the differences among categories are not statistically significant. About half of the increase or decrease in signal power can be attributed to the call destination trend for loop current, which is plotted at the bottom of

Table IV—Bell and customer-provided equipment speech signal power results

Transmission Characteristic	Bell Business				CPE Business			
	Mean	90% C.I.	Std. Dev.	Sample	Mean	90% C.I.	Std. Dev.	Sample
Near-end EPL (dBm)	−10.4	±0.7	4.6	2857	−12.5	±1.2	5.1	2552
Far-end EPL (dBm)	−12.8	±0.5	5.2	2404	−14.0	±0.5	5.1	2228
Near-end average power (dBm)	−24.9	±0.7	5.2	2857	−27.5	±1.3	5.4	2552
Far-end average power (dBm)	−28.4	±0.5	5.5	2404	−29.3	±0.4	5.1	2228
Near minus far-end EPL (dB)	2.5	±0.7	6.0	2304	1.8	±1.3	6.2	2065
Near minus far-end average power (dB)	3.4	±0.7	6.7	2304	1.6	±1.6	6.6	2065
Near-end EPL minus average power (dB)	14.6	±0.2	2.1	2857	15.1	±0.1	2.4	2552
Far-end EPL minus average power (dB)	15.7	±0.1	2.1	2404	15.3	±0.2	2.0	2228
Near-end loop current (mA)	45.1	±2.9	14.0	2957	37.8	±3.2	11.3	2715

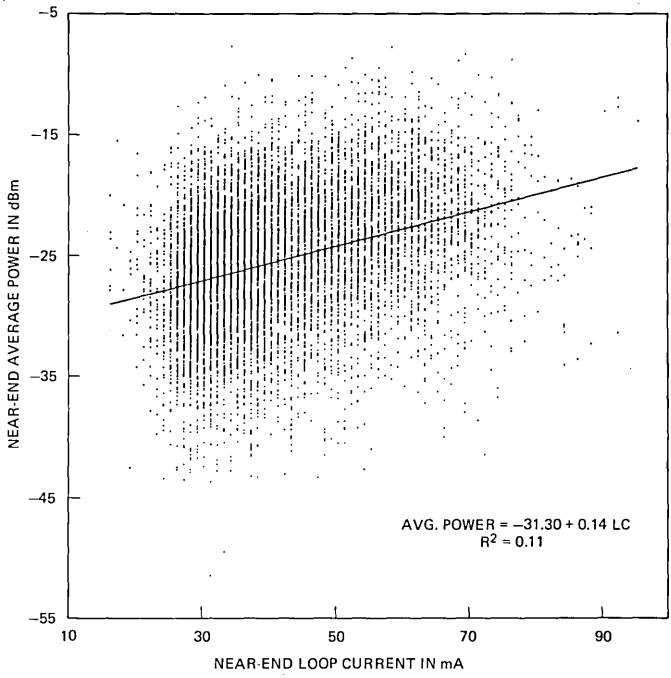


Fig. 11—Linear regression of average power on loop current.

Fig. 15. The correlation coefficients for near-end EPL and loop current are between 0.30 and 0.40 for all four destination categories. As the loop current decreases or increases, the EPL and average power trend lines follow. Since loop currents tend to be higher for business-originated calls (Section 2.3.1) and since the HNPA and FNPA categories of calls have increasingly more business-originated traffic (intrabuilding: 34 percent, interbuilding: 50 percent, HNPA: 59 percent, and FNPA: 69 percent), loop current tends to increase as the call destination becomes more remote. Interbuilding local calls, however, present an exception to this behavior which is not understood. It may be a real deviation from the overall trend, or it may be a random statistical phenomenon. Since the trends are so slight, further investigation of the interbuilding results is not warranted.

Examination of the near-end EPL and average power distributions within the individual call destination categories shows that they are close to normal in all categories except the FNPA category. In the FNPA category, both distributions modestly deviate from normality in the upper 10-percent tail due to a truncation of EPL around 0 dBm and a truncation of average power around −15 dBm. The reason for this truncation is not known; however, it represents a threshold above which speakers rarely drift. In the other call destination categories, 0 and −15 dBm signal

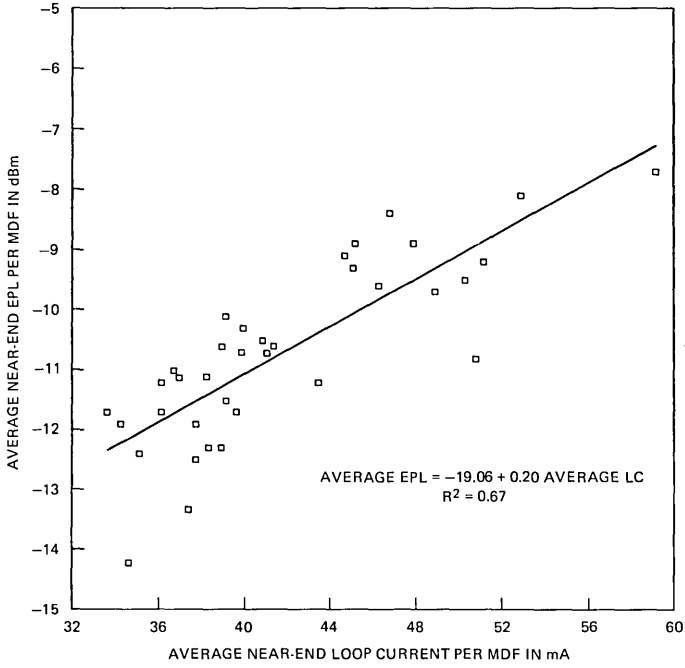


Fig. 12—Linear regression of equivalent peak level on loop current using MDF averages.

powers fall in the highest 1 percent of the EPL and average powers, respectively. The distributions for far-end EPL and average power are essentially normal in all categories.

The far-end signal powers tend to decrease as the call destination becomes more remote from the originating office due to increases in end-office-to-end-office network transmission loss. In the case of intrabuilding local calls where both parties are served by the same local switching office, the only additional network loss encountered by far-end signals is the switching office loss itself. As a result, the near- and far-end signal powers differ only slightly for intrabuilding local calls. These differences increase for interbuilding local calls and HNPA calls, which have similar far-end signal powers, due to an increase in the number of switching offices and trunks involved in the transmission path and the via net loss design[6] adopted for these arrangements of facilities. Likewise, an even greater difference between near- and far-end signal power is observed in the FNPA category. The detailed statistics associated with the trends illustrated in Fig. 15 are listed in Table V.

The correlation between near- and far-end signal powers also appears to depend upon call destination. A comparison of near- and far-end EPL provides correlation coefficients of 0.36, 0.27, 0.23, and 0.14 for intrabuilding, interbuilding, HNPA, and FNPA calls, respectively. The cor-

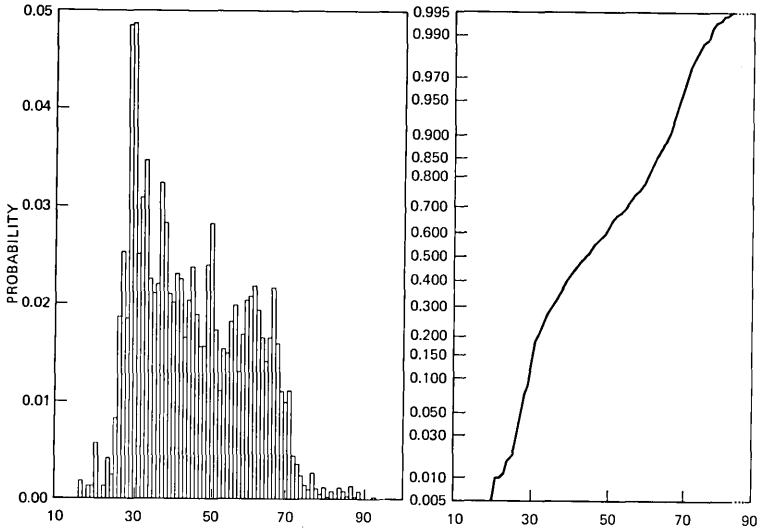


Fig. 13—Near-end loop current (mA) distribution for business.

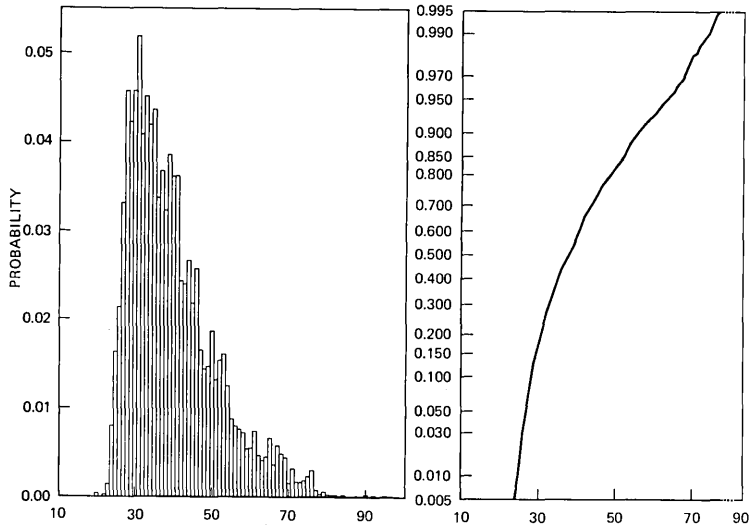


Fig. 14—Near-end loop current (mA) distribution for residential.

relation becomes poorer as the call destination becomes more remote because of the overall increasing and opposite impacts of network transmission loss and loop current on far-end and near-end signal powers, respectively.

The intrabuilding and interbuilding local call data were pooled to obtain overall local results, and the HNPA and FNPA data were pooled

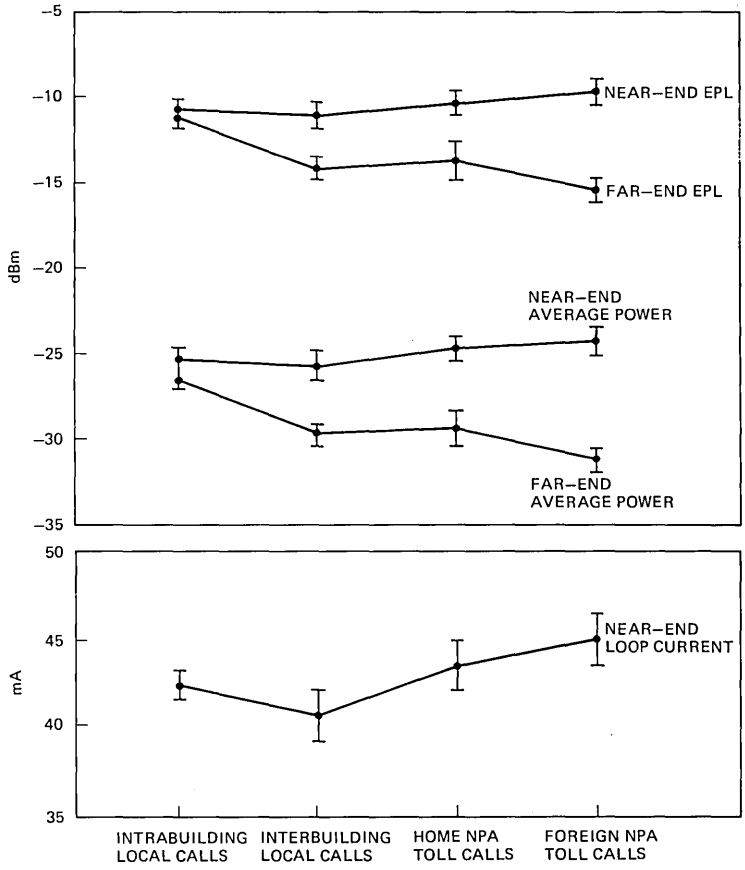


Fig. 15—Relationship of call destination to signal power and loop current.

to obtain overall toll results. Briefly, near-end toll signal powers are slightly, but not significantly, higher than near-end signal powers for local calls, and far-end toll signal powers are significantly lower than far-end powers for local calls. The reasons for these characteristics are discussed above. The only additional observation at this point is that the local loop current distribution resembles the residential distribution in Fig. 14 and the toll loop current distribution resembles the business distribution in Fig. 13. The dominance of residential and business traffic for local and toll calls, respectively, is responsible for these similarities.

2.5 Additional speech signal power analyses

The signal power data were also analyzed to determine the impact of call distance, local switch type, and several demographic factors upon

Table V—Call destination speech signal power results

Transmission Characteristic	Intrabuilding Local Calls				Interbuilding Local Calls				Home NPA Toll Calls				Foreign NPA Toll Calls			
	Mean	90% C.I.	Std. Dev.	Sample	Mean	90% C.I.	Std. Dev.	Sample	Mean	90% C.I.	Std. Dev.	Sample	Mean	90% C.I.	Std. Dev.	Sample
Near-end EPL (dBm)	−10.7	±0.5	4.6	3697	−11.0	±0.7	4.7	3704	−10.3	±0.6	4.5	995	−9.8	±0.7	4.3	763
Far-end EPL (dBm)	−11.0	±0.5	4.8	3348	−14.1	±0.4	4.8	3140	−13.6	±1.1	5.5	862	−15.4	±0.6	4.7	645
Near-end average power (dBm)	−25.3	±0.5	5.3	3697	−25.7	±0.8	5.5	3704	−24.7	±0.7	5.2	995	−24.2	±0.8	4.8	763
Far-end average power (dBm)	−26.5	±0.5	5.4	3348	−29.7	±0.5	5.2	3140	−29.3	±1.0	5.7	862	−31.1	±0.6	4.9	645
Near minus far-end EPL (dB)	0.2	±0.3	5.3	3170	3.2	±0.6	5.8	2980	3.6	±0.9	5.8	828	5.8	±1.0	6.1	617
Near minus far-end average power (dB)	1.0	±0.3	6.2	3170	3.9	±0.6	6.6	2980	4.9	±0.7	6.5	828	7.0	±1.1	6.7	617
Near-end EPL minus average power (dB)	14.6	±0.1	2.1	3697	14.7	±0.1	2.1	3704	14.4	±0.2	1.9	995	14.4	±0.3	1.9	763
Far-end EPL minus average power (dB)	15.6	±0.1	2.1	3348	15.6	±0.1	2.1	3140	15.7	±0.3	2.0	862	15.7	±0.3	2.1	645
Near-end loop current (mA)	42.4	±1.7	12.4	3875	40.7	±3.0	12.4	3864	43.6	±2.9	13.6	1029	45.1	±3.0	13.8	791

speech signal power. Call distance is defined as the airline distance between the originating and terminating local switching machines. Near-end signal power and loop current do not appear to be correlated with call distance. Far-end signal power is weakly correlated with call distance in a negative sense, due to the increase in network transmission loss which accompanies longer call distances as a result of the via net loss design.[6]

In the second of these analyses, the data were classified by originating local switching machine type. No significant relationship was found between machine type and near-end signal power.

Three demographic factors were considered in the third analysis. The first factor, geographical location, does not play an important role in determining speech signal power. While the average near-end signal power is highest in the northeast section of the country and lowest in the southwest, the range of the differences is only 2.7 dB, and the correlation between loop current and signal power accounts for about 40 percent of the difference between geographic areas. The second factor, city or town population, tends to mask rather than uncover relationships between signal power and population. A more appropriate measure is the population density of the exchange served by the local telephone office. The third demographic factor, locality type, was defined to capture the impact of population density upon speech signal power. Five locality types were considered: downtown areas of large and midsize cities, downtown areas of small towns, outer-urban areas, and suburban areas. Large cities were defined as cities with populations of 100,000 or more people; mid-size cities were defined as cities with populations ranging from 20,000 to 100,000 people; and small towns were defined as cities or towns with populations of 20,000 or less people. The outer-urban classification denotes areas with a mixture of residential dwellings and business establishments on the outlying fringes of large cities, and the suburban classification denotes areas which primarily contain residential dwellings. The average near-end EPL and loop current both exhibit the same trends with locality types. Both are highest for downtown MDFs in large cities and lowest for outer-urban areas. These results correlate with the fact that in the first case the population of customers is rather concentrated, and they tend to have relatively short loops, while in the second case the population of customers is rather widespread, and they tend to have relatively long loops. Between these extremes, the average EPL and loop current for small towns are higher than for mid-size cities, and both have higher averages than suburban areas. As illustrated in Figs. 16 and 17, the differences among the categories are not large; however, they do suggest a dependence of loop current and, as a result, EPL upon varying densities of populations.

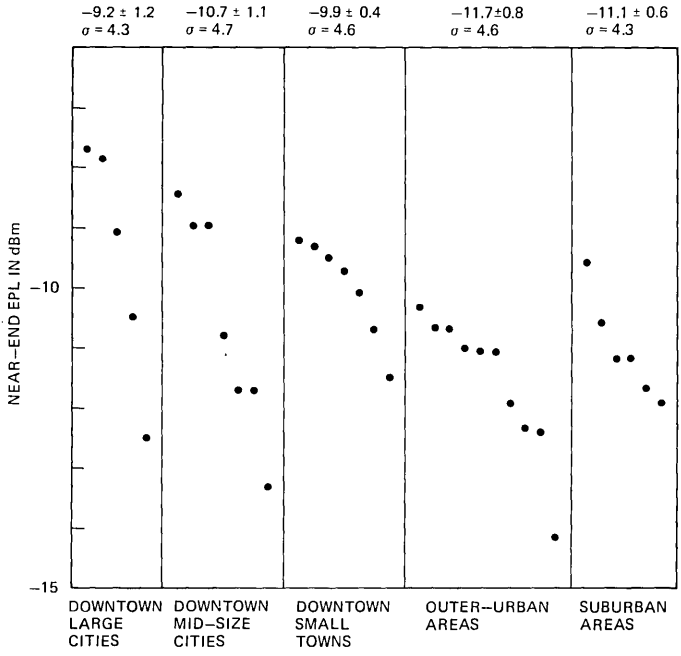


Fig. 16—Average signal power per MDF by locality type.

III. METHODOLOGY

3.1 Statistical survey sample plan

The Loop Signal Power Survey sampling plan consists of three major components—a precise definition of the target population and parameters, a scheme for the selection and measurement of a sample of calls, and the choice of the estimation formulas. Section 3.1.1 defines the target population and parameters, Section 3.1.2 describes the scheme used to select and measure a statistical sample of calls, and Section 3.1.3 describes the statistical estimation and confidence interval formulas used to estimate the target parameters.

3.1.1 Target population and measured parameter definitions

The target population consists of voice calls originating over the public switched network where the subscriber's loop is classified as business, single party residence, coin semipublic, Private Branch Exchange (PBX), or Centralized Exchange (centrex) service. The aggregate of subscriber loops in the target population are naturally partitioned according to the local MDF in which they terminate. In addition, the subscriber loops terminating in an MDF are naturally dichotomized into a customer-provided equipment (CPE) substratum and a Bell equipment substratum. A loop was identified as belonging to the CPE substratum when

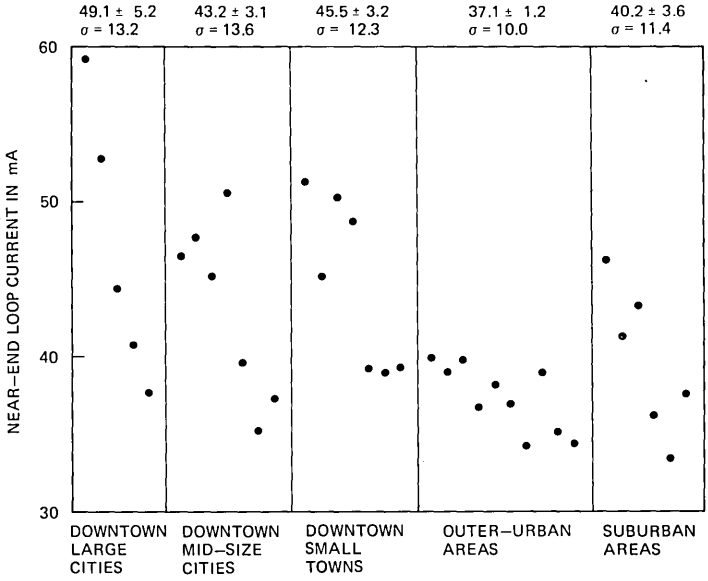


Fig. 17—Average loop current per MDF by locality type.

the local operating company billing records and a follow-up station verification identified the connection to the subscriber's loop of a protective connecting arrangement (PCA) listed in Table VI. A PCA is designed to interconnect non-Bell terminal equipment with the Bell System public switched network.

For potential statistical advantage, the MDFs were partitioned into 12 strata according to the average 1970 population census of the communities within the plant district where an MDF was located. The 12 strata were constructed so that they are approximately the same size with respect to the total number of business, residence, PBX, centrex, coin semipublic, and switched data telephone lines terminating on MDFs within the stratum. This form of stratification was suggested by the results of the 1960 Speech Volume Survey, which indicated a correlation of speech volume with city population. Stratification by city size offered the potential for reduction of the variability in speech signal power within

Table VI—Protective connecting arrangements (PCA)

PCA USOC*	Associated Non-Bell Terminal Equipment
STP	Key telephone system
STC	Single line set
C2ACP	Single line or key telephone system
CD8, CDH	PBX or centrex CU
CDA, CD1, CD7, CD9	Cord switchboard or console

* USOC—Bell System Universal Service Order Code.

each of the strata and, as a result, an increase in the precision of estimates of the mean signal power.

In general, the choice of the criterion for stratification is arbitrary and does not affect the validity of the final survey conclusions; however, a judicious choice of a stratification scheme can lead to an estimate of the mean with a smaller confidence interval than would be obtained otherwise.

Each loop associated with the target population is indexed by its stratum number, MDF number within a stratum, substratum number (e.g., 1 Bell, 2 CPE) and loop number within an MDF substratum.

The target population parameters estimated in the Loop Signal Power Survey are defined by the ratio

$$R = Y/X,$$

where

$$Y = \sum_{h=1}^{L} \sum_{i=1}^{N_h} \sum_{a=1}^{D_{hi}} \sum_{j=1}^{M_{hia}} \sum_{k=1}^{Q_{hiaj}} Y_{hiajk},$$

X is defined similarly to Y with Y_{hiajk} replaced by X_{hiajk},

N_h is the number of MDFs located in class 5 offices in stratum h, for $h = 1, 2, \ldots, L$,

D_{hi} is the number of substrata into which the subscriber loops that terminate in the ith MDF of stratum h are partitioned ($D_{hi} = 2$),

M_{hia} is the number of subscriber loops that are in substratum a and terminate on the ith MDF in stratum h,

and

Y_{hiajk} and X_{hiajk}, $k = 1, 2, \ldots, Q_{hiaj}$, represent measurements associated with the Q_{hiaj} completed calls which originate from loop $(hiaj)$. Loop $(hiaj)$ is identified as the jth loop terminating in substratum a of the ith MDF in stratum h.

Some examples of applications of the ratio parameter R are given below.

Application One: Fraction of Calls Where the Mean Transmitted Signal Power Exceeds Some Threshold

Suppose Y_{hiajk} is defined as 1 if the kth completed call on loop $(hiaj)$ is in the target population and the mean signal power exceeds some threshold T, and 0 otherwise. Second, suppose X_{hiajk} is defined as 1 if this call is in the target population, and 0 otherwise. R is then equal to the fraction of completed calls in the target population for which the transmitted mean signal power exceeds T. This form of the ratio parameter is applicable to target populations such as completed calls (toll and/or local) originating from the Bell and/or CPE subclasses of loops.

Application Two: The Mean Originating Signal Power Per Call

Suppose X_{hiajk} is defined as in Application One, and Y_{hiajk} is defined as a measure of signal power of the kth completed call originating on loop $(hiaj)$, then R is equal to the mean originating signal power per call.

3.1.2 Survey sampling scheme

The calls which were measured in the Loop Signal Power Survey were statistically selected in such a way as to permit precise estimates of the population parameters described in Section 3.1.1 and at the same time limit the costs of obtaining the measurements. The actual statistical sample selection scheme used was a classical three-stage sampling scheme with stratification and substratification. From each of the 12 strata described in Section 3.1.1, three MDFs were selected with probabilities of selection proportional to estimates of the total number of business, residence, PBX, centrex, and coin semipublic lines terminating on each MDF. The locations of the 36 sampled MDFs are illustrated in Fig. 18. A stratified random sample of CPE and Bell loops, which terminated on the 36 MDFs, was selected, specially designed measurement equipment was connected to these sampled loops, and signal power measurements were made on a sample of calls originating over the loops. The selection of the CPE loops was made from a billing records inventory of subscriber telephone numbers that were being billed for a PCA with one of the Universal Service Order Codes (USOC) listed in Table VI. A random sample of Bell loops was obtained by generating a list of random four-digit numbers and prefixing a local three-digit NNX code for each NNX associated with the MDF. These lists were forwarded to the local repair service bureau for determination of the class of service of each

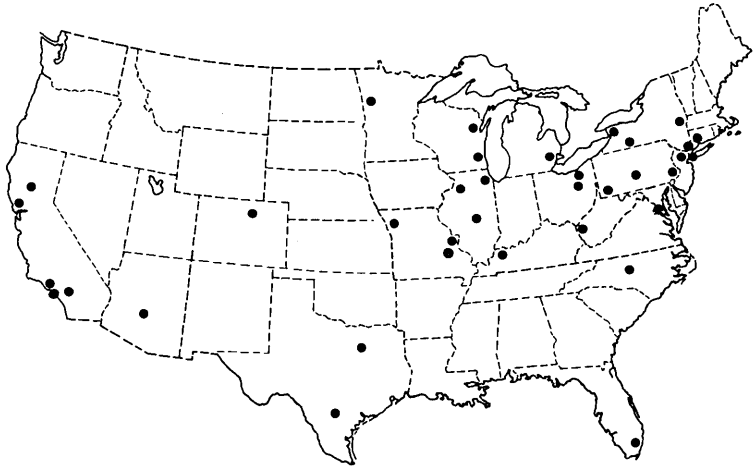


Fig. 18—Locations of sampled MDFs.

telephone number and the location of the loop on the MDF. A stratified random sample of CPE and Bell loops, identified as members of the target population, was ordered according to the location on the vertical side of the MDF. Approximately 1 week prior to the scheduled arrival of the Bell Laboratories survey team, a verification was made by local operating company craft people to assure that each selected line was working, that the telephone number-cable-pair and horizontal frame assignments were correct, and that no bridged lines were present. From this verified list, a stratified sample of up to 30 CPE loops and at least 69 Bell loops (for a total of 99) were selected for connection to the survey equipment. The equipment included a device which, when activated, scanned the 99 loops for an originating off-hook signal. Following seizure of the loop and the establishment of a connection, the measurement process was started manually if a conversation ensued. Conversation was detected by utilizing an equipment operator's monitor channel which provided unintelligible speech during periods of conversation through the use of a low speech sampling rate. Because toll calls were relatively scarce, provision was made for the equipment operator to abort the measurement of local calls to obtain additional toll calls. The measurement period in a local office was 3 days.

The survey equipment provided peg count data from which the number of originated completed calls was estimated for each loop. These data formed the basis for traffic weights used to estimate the target population parameters.

3.1.3 Estimation formulas and confidence intervals

This section is devoted to a discussion of the statistical estimation formulas that are used to estimate the ratio parameter R. These formulas are tailored to the survey sample design discussed in Section 3.1.2. The form of the estimation formulas require the following information relative to the sampling plan:

n_h—the number of sampled MDFs in primary stratum h for $h = 1, 2, \ldots, L$. ($n_h = 3$ for $h = 1, 2, \ldots 12$).

z_{hi}—the probability of selection into the first stage sample of the ith sampled MDF in stratum h for $i = 1, 2, \ldots, n_h$ and $h = 1, 2, \ldots, L$.

m_{hia}—the number of measured subscriber loops that belong to the ath substratum of the ith sampled MDF in stratum h for $i = 1, 2, \ldots n_h$; $a = 1, 2, \ldots D_{hi}$, and $h = 1, 2, \ldots, L$.

q_{hiaj}—the number of calls associated with loop ($hiaj$) on which signal power measurements were made.

L, M_{hia}, D_{hi} and Q_{hiaj} are defined as in Section 3.1.1, and

(x_{hiajk}, y_{hiajk}), $k = 1, 2, \ldots, q_{hiaj}$ represents a sample of q_{hiaj} values of (X_{hiajk}, Y_{hiajk}), $k = 1, 2, \ldots Q_{hiaj}$, where

X_{hiajk} and Y_{hiajk} are defined as in the definition of R.

A three-stage estimator of the ratio $R = Y/X$ where Y and X are defined as in Section 3.1.1 is

$$r = y/x,$$

where

$$y = \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hia}} \sum_{k=1}^{q_{hiaj}} w_{hiaj} y_{hiajk},$$

$$w_{hiaj} = \frac{1}{n_h} \frac{1}{z_{hi}} \frac{M_{hia}}{m_{hia}} \frac{Q_{hiaj}}{q_{hiaj}},$$

and x is defined similarly to y with y_{hiajk} replaced by x_{hiajk}.

The mean squared error of r is defined as

$$\text{VAR}(r) = E(r - R)^2,$$

where $E(\cdot)$ denotes expected value.

A consistent estimator of VAR(r) is

$$v(r) = \frac{1}{x^2} \sum_{h=1}^{L} \frac{1}{n_h(n_h - 1)} \sum_{i=1}^{n_h} \left[\frac{y_{hi} - r x_{hi}}{z_{hi}} - \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{y_{hi} - r x_{hi}}{z_{hi}} \right]^2,$$

where

$$y_{hi} = \sum_{a=1}^{D_{hi}} \sum_{j=1}^{m_{hia}} \sum_{k=1}^{q_{hiaj}} \frac{M_{hia}}{m_{hia}} \frac{Q_{hiaj}}{q_{hiaj}} y_{hiajk}$$

and x_{hi} is defined similarly to y_{hi} with y_{hiajk} replaced by x_{hiajk}.

An application of the Central Limit Theorem yields an approximate 90-percent confidence interval for R as the interval

$$(r - 1.645\sqrt{v(r)}\,,\ r + 1.645\sqrt{v(r)}).$$

3.2 Data acquisition plan

In this section, requirements pertaining to acquisition equipment capacity, compatibility, transparency, privacy, etc., are summarized, and a block diagram of the Loop Signal Power Survey acquisition equipment is discussed.

3.2.1 Requirements

As indicated in Section 3.1, the sample plan called for access to 99 customer loops in each of 36 class 5 offices and measurements of near- and far-end signal power on live calls. Determination of call destination required the detection of call originations on loop start and ground start lines, and the detection of dial pulse and *TOUCH-TONE*® address information. Because of the loop-to-loop and call-to-call variability in impedance at the MDF interface, the measurement of real power was required rather than bridged voltage. In the course of accessing and

measuring calls, no detectable impairment (loss or switching clicks) was to be added to the connection. Monitoring of intelligible speech was prohibited by privacy considerations. Speech signals are predominantly half-duplex in nature; however, both parties sometimes talked at the same time. Because the point of measurement was a two-wire point, it was necessary to devise a method to sort the speech signal data into two categories, near-end and far-end.

3.2.2 Data acquisition equipment

Figure 19 is a block diagram of the equipment used to acquire speech signal power data. The 99 customer loops were accessed at the protector socket of the MDF. Access cables connected the customers' loops to the acquisition console protector panel. This panel provided series access to 99 loops, circuit protection, and an electrical interface with the instrumentation switch. This interface contained current sensing resistors for the detection of metallic speech current and loop dc current. Modified service observing equipment was bridged across the tip-ring interface at this point to allow the detection of outgoing call seizures and the de-

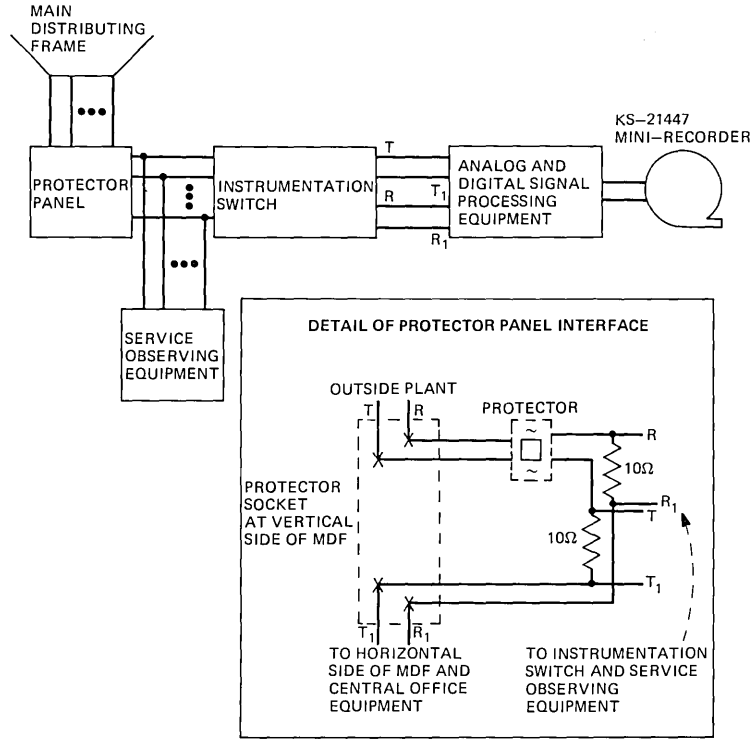


Fig. 19—Loop signal power survey data acquisition console.

tection of dial pulse/*TOUCH-TONE* address digits. The instrumentation switch connected the four leads associated with the current sensing resistors of one of the 99 loops to the analog signal processing equipment for the detection, amplification, and filtering of the metallic speech voltage and current.

The resulting voltage and current signals were simultaneously sampled at the rate of 200 samples per second using dual 12-bit A-D converters. The sampled data were stored in a buffer memory, combined with label information, and written in 16-kb blocks on a minirecorder magnetic tape unit. A paper-tape printer recorded off-hook event times for each of the 99 loops so that traffic weights referred to in Section 3.1 could be determined. In addition, the dialed area and office code were recorded on the tape. The digitally recorded speech signal data were subsequently analyzed in a manner described in the next section.

The loss due to the current sensing resistors and bridged equipment was negligible. This, combined with click supression circuitry, made the measurement equipment transparent from the customer's point of view. The low rate of sampling made the recorded speech signal unintelligible but allowed the recovery of pertinent signal power information. A low speech sampling rate was also used to make the equipment operator's monitor channel unintelligible, yet permit the identification of call progress signals. The acquisition of simultaneous speech voltage and current samples permitted the discrimination of the near-end from the far-end talker in a manner discussed in the next section.

3.3 Analysis of data

This section explains how voltage and current samples were processed to obtain measures of speech signal power for each talker in the two-way conversations.

3.3.1 Raw speech signal power data processing

The raw data upon which speech signal equivalent peak level (EPL) and average power estimates are based consisted of metallic speech voltage and current samples. The metallic speech voltage and current on the loop were amplified and filtered to exclude signals higher than 4 KHz and remove the effect of 60 Hz, its first two odd harmonics, and low frequency noise below 100 Hz. The resultant voltage and current analog signals were then simultaneously sampled at the rate of 200 samples per second using two 12-bit linear A-D converters. The digital sampled data were then recorded on tape cartridges, which were later reformatted onto standard computer tape.

The first step in computer processing of the digitally recorded signals consisted of removal of dc bias produced in the analog signal processing filters and computation of the instantaneous power (watts) associated

with each voltage-current sample pair. The equipment was designed so that the power values were positive (voltage and current in phase) when the signal source was the near-end talker and and negative (voltage and current out of phase) when the signal source was the far-end talker.

3.3.2 Discrimination of near-end and far-end talkers

Conversational speech is predominantly half-duplex, but brief periods occur when both talkers are active at the same time. The stream of instantaneous power samples is therefore positive or negative for half-duplex talk-spurts. However, during double talking, the sign of the power samples may change rapidly and the magnitudes of the power samples become useless for estimation of near-end or far-end talker power. To properly sort the power sample stream into two distinct "bins" corresponding to the near-end and far-end talkers, empirical algorithms were developed in laboratory simulations, and one algorithm (SGN algorithm) was chosen for use during the speech signal processing phase of the survey.

The SGN algorithm uses the sign and magnitude of the power in short subsequences of the stream of speech power samples to generate two sequences of speech power samples corresponding to near-end and far-end talkers.

Let $\{p\}$ be the sequence of instantaneous speech signal power values computed from the relationship: $p = v \cdot i$, where $\{v\}$ and $\{i\}$ are sequences of instantaneous, simultaneous samples of speech signal metallic voltage and current, respectively.

Let the sequence $\{p\}$ be divided into consecutive subsequences of length l. Associated with the ith subsequence is the average power:

$$\overline{p}_i = \frac{1}{l} \sum_{k=il-l+1}^{il} p_k.$$

$$\text{Let } \mathrm{SGN}(\overline{p}_i) = \begin{cases} -1 \text{ if } \overline{p}_i > 0 \\ 0 \text{ if } \overline{p}_i = 0 \\ +1 \text{ if } \overline{p}_i < 0. \end{cases}$$

The SGN algorithm depends on two conditions for every subsequence:

Condition 1: $\mathrm{SGN}(\overline{p}_i) = \mathrm{SGN}(\overline{p}_{i-1})$

Condition 2: $|\overline{p}_i| \geq \alpha |\overline{p}_{i-1}|$.

If either condition is true, then SGN $(\overline{p}_i)$ determines the sources of the speech signal for the ith subsequence. As stated earlier, the sign convention is such that a positive value indicates that the near-end talker is the source (far-end samples set to 0), and a negative value indicates that the far-end talker is the source (near-end samples set to 0). After the source is determined, the nonzero power samples are set positive and placed in the appropriate (near- or far-end) sequence.

If neither of the above conditions is true, then the direction is indeterminant and all power samples in the ith subsequence are set to 0. Laboratory investigations established that the values $l = 2$ and $\alpha = 10$ give good performance with the sample rate used in the survey (200 samples per second). The output from the SGN algorithm consists of two sequences of positive instantaneous signal power samples representing the near-end and far-end talkers.

3.3.3 Measures of speech signal power

Two measures of speech signal power are developed from each of the near-end and far-end sequences described above. The first measure is the average speech signal power defined over the observation interval (generally about a minute) as follows:

$$\text{Near-end average power} = 30 + 10 \log \frac{1}{n} \sum_{k=1}^{n} p_k \text{ near-end (dBm)}$$

$$\text{Far-end average power} = 30 + 10 \log \frac{1}{n} \sum_{k=1}^{n} p_k \text{ far-end (dBm)},$$

where p_k-end represents the elements in the sequence of instantaneous power samples for the direction of interest, and n is the total length of the power sample sequence.

The second measure used to characterize speech signal power is an estimate of the peak power in the distribution of samples of talker signal power. The estimator is the empirical equivalent peak level (EPL), developed by Brady. A complete discussion of the EPL and its properties is given by Brady in Ref. 2. The EPL is developed from the power sample sequence for the direction of interest as follows.

Let the instantaneous power of the kth sample be defined as:

$$p_k = v_k i_k \text{ watts.}$$

In logarithmic units,

$$p_k = 10 \log p_k \text{ (dBw).}$$

Define a threshold ϕ and multiplier δ_k so that:

$$\delta_k = \begin{cases} 1 & \text{if } p_k > \phi \\ 0 & \text{otherwise} \end{cases}.$$

The average power over threshold is defined:

$$\overline{p}_\phi = 10 \log \left(\frac{\sum_{k=1}^{n} p_k \delta_k}{\sum_{k=1}^{n} \delta_k} \right).$$

Now define $D = \overline{p}_\phi - \phi$ dB. From D compute Δ using the following empirical rule:

$$D \leq 6.75, \text{ then } \Delta = (D - 2.75)/0.4$$

$$6.75 < D \leq 13.5, \text{ then } \Delta = D/0.675$$

$$13.5 < D, \text{ then } \Delta = (D + 2.88)/0.819.$$

From Δ compute EPL as:

$$\text{EPL} = \Delta + \phi.$$

Some important properties of the EPL are that it is independent of the talker's activity since it is not affected by the silent periods in the conversation, and its estimate varies little over a wide range of threshold values. Some laboratory investigations indicate that a threshold of 10 to 20 dB below EPL gives good performance in the presence of noise; a threshold of 20 dB below EPL was selected as giving the best noise rejection without discarding an excessive number of samples. The EPL computation was iterated until the threshold was 20 ± 3 dB below the EPL value.

IV. COMPARISON WITH PREVIOUS DATA

In 1960, measurements of talker volume were made on live traffic using VU meters.[1] These measurements of talker volume are compared with the current survey results, which have been translated from EPL to VU using an empirical correction factor. These results are listed in Table VII together with the 1960 survey results.

The 1960 survey results differ substantially from the current results in that the toll volumes were substantially higher in 1960 and the ranges of volumes within the various call destination categories were substantially greater. There have been some substantial changes in the telephone plant since 1960 that may help to explain these differences. The proportion of toll grade battery has decreased substantially, resulting in a decrease in toll call speech volume. Loss plan improvements, the phasing out of the 300-type telephone set, and the growth of direct trunking have all tended to increase the uniformity of service in the network and make it more transparent to customers. The apparent result is a network with remarkable uniformity of speech signal power.

Table VII—Comparison with 1960 speech volume survey

	1960		1975–1976	
Call Destination	Average VU	Std. Dev.	Average VU	Std. Dev.
Intra-building	−24.8	7.3	−22.2	4.6
Inter-building	−23.1	7.3	−22.5	4.7
Toll	−16.8	6.4	−21.6	4.5

V. ACKNOWLEDGMENTS

Many individuals contributed to the Loop Signal Power Survey. P. W. Freeman, J. M. MacMaster, and E. J. Vlacich designed and assembled the test equipment, T. W. Thatcher, Jr. handled all phases of logistic support, and F. Grizmala, S. Vitale, and P. R. Wild developed software packages and processed the data with the help of C. Keinath. In addition, many individuals from Bell Laboratories and the Bell System operating companies participated in the data collection phase of the survey. We extend our thanks to all who have helped in this project.

REFERENCES

1. K. L. McAdoo, "Speech Volumes on Bell System Message Circuits—1960 Survey," B.S.T.J., *42,* No. 5 (September 1963).
2. P. T. Brady, "Equivalent Peak Level: A Threshold-Independent Speech-Level Measure," J. Acoust. Soc. Am., *44,* No. 3 (September 1968).
3. Bell System Center for Technical Education, *Telecommunications Transmission Engineering,* Vol. 1, Chapter 12, 1974.
4. D. H. Merchant, unpublished works.
5. P. A. Gresh, "Physical and Transmission Characteristics of Customer Loop Plant," B.S.T.J., *48,* No. 10 (December 1969).
6. *Notes on Distance Dialing,* Section 6, New York: American Telephone and Telegraph Company, 1975.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 7, September 1978
Printed in U.S.A.

An Adaptive PCM System Designed for Noisy Channels and Digital Implementations*

By DEBASIS MITRA and B. GOTZ

(Manuscript received November 29, 1977)

We propose a new adaptive quantization scheme for digitally implementing PCM and DPCM structures. The arithmetics we develop for the digital processing are useful as well in the implementation of previously existing schemes for adaptive quantization. Two objectives are stressed here: (i) The system must be robust in the presence of noise in the transmission channel which causes the synchronization between quantizer adaptations in the transmitter and receiver to deteriorate. (ii) It must also minimize the complexity of the digital realization. In addition to the above objectives, we require, of course, good fidelity of the processed speech waveform. The problem of synchronization in digital implementations where the constraint of finite precision arithmetic exists has not been addressed previously. We begin by examining an existing, idealized adaptation algorithm which contains a leakage parameter for the purpose of deriving robustness. We prove that, to provide the necessary synchronization capability without impairing the quality of speech reproduction, it is necessary to use a minimum, unexpectedly large, number of bits in the machine words and, additionally, to carefully specify the internal arithmetic, as is done here.

The new scheme that we propose here uses an order of magnitude less memory in an ROM-based implementation. The key innovations responsible for the improvement are: (i) modification of the adaptation algorithm to one where leakage is interleaved infrequently but at regular intervals into the adaptation recursion; (ii) a specification of the internal machine arithmetic that guarantees synchronization in the presence of channel errors. A detailed theoretical analysis of the statistical behavior of the proposed system for random inputs is given here. Results of a simulation of a realistic 16-level adaptive quantizer are reported.

* A short version of this paper was presented at the International Conference on Communications, Toronto, June 1978.

I. INTRODUCTION

We propose a new scheme for adaptive quantization which is particularly well suited to the digital implementation of PCM and DPCM structures. In the course of this work, we have developed arithmetics for the digital processing that are useful as well in the implementation of previously existing schemes for robust quantization.

The exacting requirements on adaptive quantization stemming from the broad dynamic range and rapid transient behavior of speech are well known. Two additional objectives are given equal importance here: (*i*) To make the system robust in the presence of channel errors. Thus, while channel errors may cause the quantizer adaptations in transmitter and receiver to be put out of synchronization,* a mechanism must exist which acts to rapidly restore the synchronization during periods of error-free transmission. (*ii*) To minimize the complexity of the digital realization; specifically, to minimize the length of the internal words in the digital processors and to facilitate the multiplexing of the hardware.

Systems do exist in the literature for robust quantization in the presence of noisy channels; one such system is described below in some detail. However, the problem of synchronizing the quantizer adaptations in the transmitter and receiver in digital implementations, where the constraint of finite precision arithmetic exists, has not been addressed previously. We prove that, to provide the necessary synchronization capability without impairing the quality of speech reproduction, it is necessary to use an unexpectedly large number of bits in the internal words of the digital processors at both sites and, additionally, to carefully specify the internal arithmetic (which we do). If the digital processing is implemented using ROMs, as is being proposed, the long internal word length is reflected in large memory requirements and therefore costly implementations as well as exposure to new errors in the processing.

The scheme that we propose here uses an order-of-magnitude less memory in an ROM-based implementation in both the transmitter and receiver. This is for comparable performance with respect to loading characteristic, signal-to-noise ratio, and the synchronization capability. Another advantage not reflected in the above estimate is the fact that the essential costly digital component, the ROM, as distinct from other less costly components such as adders, is used only for a small fraction of the total operating time. Thus, further economies may be effected through multiplexing the ROM. The key innovations are: (*i*) the modification of the adaptation algorithm which allows the internal word length of the digital processors to be reduced significantly; and (*ii*) a specification of the internal arithmetic that guarantees synchronization in the presence of channel errors. As mentioned previously, the arithmetic is also applicable in digital implementations of previously existing adaptation algorithms.

* In our usage, synchronization is synonymous with tracking.

A byproduct of the work reported here is that it establishes a link between two hitherto unconnected areas, namely, finite-arithmetic digital signal processing and waveform quantization in the presence of a noisy channel. The problem of synchronizing two geographically separated digital processors gives rise to quite novel requirements on the processing, and we expect that the problem will be a subject of further investigation in the future.

The paper is organized as follows. In Section 1.1 we describe an existing quantizer adaptation scheme and the associated synchronization problem. Section II is devoted to the basic description of the new scheme. Section 2.1 introduces the key idea underlying the scheme. Section 2.2 considers the digital implementation of the system, and Section 2.3 considers the synchronization behavior of the resulting system. Section III is devoted to the probabilistic analysis of the behavior of the proposed algorithm. The basic notions of the bias functions, central log step sizes, and load curves are introduced, and the qualitative results proved in their connection are stated. In Section IV, some computational results are presented in the context of a realistic 16-level quantizer that has been proposed and investigated previously in connection with an industrial application. We try to illuminate the topics considered in Sections II and III through examples involving this particular quantizer. Four appendices to the paper present the detailed technical derivations.

On account of the length of the paper, we considered it desirable to include a final section, Section V, which summarizes and puts into perspective the key results obtained in the preceding sections.

We should mention that the digital implementation of adaptive DPCM systems is under investigation within Bell Laboratories in connection with TASI-D, subband voice coding, and new channel banks. The work reported here is a research study and not a description of a developed design.

1.1 Background and description of the problem

We begin by describing a system proposed in Ref. 1 which, unlike earlier systems upon which it is based,[2–5] possesses the capacity to recover from past channel errors during periods of error-free transmission.

1.1.1 An existing idealized scheme for robust quantization

Let $\Delta(i)$ (see Fig. 1) denote the *step size* of a quantizer, with $2N$ levels, at the ith sampling instant; $\Delta(i)$ is adapted according to the rule

$$\Delta(i+1) = \Delta(i)^{\beta} M(i), \qquad i = 0,1,2,\cdots \tag{1}$$

where β, $0 < \beta < 1$, is the leakage constant and $M(i)$ is the *multiplier* at time i. $M(i)$ is selected from a prespecified collection of multipliers $\{M_1, M_2, \cdots, M_N\}$ according to the rule:

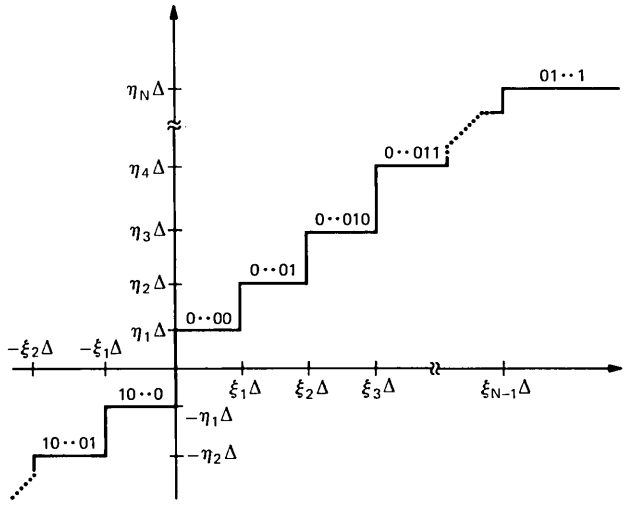


Fig. 1—The quantizer. A natural coding scheme is displayed. The step size is time-varying and the parameters (ξ_n) and (η_n) are prespecified and fixed.

$$\text{If } \xi_{r-1}\Delta(i) \leq |x(i)| < \xi_r\Delta(i), \text{ then } M(i) = M_r, \tag{2}$$

where $x(i)$ is the input signal variable (speech or data) at time i and $0 = \xi_0,\xi_1, \cdots ,\xi_{N-1},\xi_N = \infty$ are fixed, ordered parameters of the quantizer,* Fig. 1. The multipliers are also ordered, i.e.,

$$M_1 \leq M_2 \leq \cdots \leq M_N.$$

It is widely recognized[6,7] that (1) is not in a form convenient for implementation, even analog implementation. To utilize conventional multipliers, it is necessary to work with the log-transformed version of (1).

Denote the *log step size* by $d(i)$, where

$$d(i) \triangleq \log_Q \Delta(i), \tag{3}$$

Q being a fixed number greater than 1, and the *log multipliers* by

$$m(i) \triangleq \log_Q M(i),\ m_r \triangleq \log_Q M_r,\ 1 \leq r \leq N. \tag{4}$$

Also let

$$\bar{\xi}_r \triangleq \log_Q \xi_r, \qquad 1 \leq r \leq N. \tag{5}$$

Thus, from (1) and (2),

$$d(i+1) = \beta d(i) + m(i), \qquad i = 0,1,\cdots \tag{6a}$$

where

* When the parameters $\{\xi_r\}$ and $\{\eta_r\}$ are spaced equal distances apart, the quantizer is usually referred to as a uniform quantizer and it is natural to call Δ the "step size." However, for nonuniform quantizers, the term "step size" is less natural and other candidates are "scale" and "range." However, since there is no reason for confusion, we retain the familiar term "step size."

$$m(i) = m_r \text{ iff } \bar{\xi}_{r-1} + d(i) \leq \log_Q |x(i)| < \bar{\xi}_r + d(i). \tag{6b}$$

The only information that is coded and transmitted at time i is that concerning the quantizer output which uniquely determines the selected log multiplier $m(i)$. A natural coding scheme is exhibited in Fig. 1. The recursion in (6) is implemented at both the transmitter and receiver. We let $m'(i)$ denote the log multiplier corresponding to the received code word at time i, and we employ the natural notation $d'(i)$ to denote the log step size in the receiver. The reconstruction, $R(i)$, at the receiver of the input signal variable is done according to the rule:

$$\text{If } m'(i) = m_r \text{ then } |R(i)| = \eta_r Q^{d'(i)}, \tag{7}$$

where η_r, $1 \leq r \leq N$, are also prespecified, fixed parameters of the quantizer, as shown in Fig. 1. The sign of the reconstructed value is obtained from the sign bit, usually the first and shown as such in Fig. 1, in the received code word.

The synchronization capability of the system, i.e., the capability possessed by the solutions of the recursions, $\{d(\cdot)\}$ and $\{d'(\cdot)\}$, at the transmitter and receiver to approach each other during error-free transmission is entirely due to the presence of the leakage parameter β. For if $d(0)$ and $d'(0)$ are two, possibly different, initial values of the log-step sizes at the commencement of an epoch of error-free transmission, then during the epoch

$$|d(i) - d'(i)| = \beta^i |d(0) - d'(0)|, \qquad i \geq 0. \tag{8}$$

The notion of introducing leakage as a mechanism for deriving robustness in the presence of a noisy channel is a well-known one in communication practice; witness, the leaky delta-modulator.[8]

As far as the synchronization of the transmitter and receiver adaptations is concerned, eq. (8) implies that decreasing β provides improved quality. However, there is an accompanying price. The data in Fig. 5 of Ref. 1 together with the theory developed here in Sections 3.2 and 3.3 on the load curves (which describe the statistical behavior of the step size for random inputs) show that the statistical dynamic range of the step size is reduced rapidly with decreasing β, with a concomitant deterioration of the quality of the reconstruction.* Recent subjective tests[10] have shown that it is very unlikely that β less than $^{63}/_{64}$ can provide acceptable quality speech reproduction.

Herein lies the gist of the problem: For good quality reproduction, the leakage parameter must necessarily be very close to 1, and this, on the other hand, makes it difficult to provide good quality synchronization. It is thus necessary to walk a narrow path between too small leakage and too large leakage. As we see next, the constraint of finite precision

* Numerous related topics are treated analytically in Ref. 9.

arithmetic imposed by a digital implementation compounds the design problem.

1.1.2 Digital implementations

Equation (6) assumes continuous values of $d(\cdot)$ and infinite precision arithmetical operations, and hence it can only serve as an ideal in a digital implementation. An all-digital coder will have only a limited dictionary or total number (typically, ≥ 32, ≤ 128) of possible log step sizes. We will consider the log step sizes to be integers varying from 0 to $2^K - 1$; thus, typically, $5 \leq K \leq 7$. It is necessary to introduce the notion of an *internal machine word* with K integer bits and, say, F fractional bits (the need for fractional bits will become apparent shortly); the log step size is obtained from the internal machine word at time i, $y(i)$, by means of an external arithmetic, such as truncation. Although later we will consider other possibilities, for the purpose of this discussion let us assume that the external word at time i, which is the log-step size at that time, is simply the integer part of the internal word at time i, i.e.,

$$d(i) = [y(i)]_{\text{truncate}}, \qquad i = 0,1,2 \cdots. \tag{9}$$

The machine implementation of the ideal recursion in (6) is

$$y(i+1) = \langle \beta y(i) \rangle + m(i), \qquad i = 0,1,2, \cdots, \tag{10}$$

where $\langle \beta y(i) \rangle$ denotes some procedure, such as rounding, for taking $\beta y(i)$ into a $(K+F)$-bit word. It will turn out later that this operation is best viewed with greater generality as a mapping f of $(K+F)$-bit words, with F fractional bits into other such words. Thus we restate (10) as*

$$y(i+1) = f\{y(i)\} + m(i), \qquad i = 0,1,2, \cdots. \tag{10'}$$

It will be assumed that all the log multipliers $\{m_r\}$ have at most F fractional bits each, which ensures that if $y(i)$ is a $(K+F)$-bit word then so is $y(i+1)$.

Figure 2 shows an example of the most direct procedure for generating the discrete map $f(y)$, namely, by rounding βy to the nearest machine word. In the example, considered $F = 1$ so that the spacing between machine words is $2^{-F} = \frac{1}{2}$. A feature common to such maps is that segments of unit slope are juxtaposed between other segments of zero slope which we call "breaks."

If, as before, we distinguish the quantities associated with the receiver by the superscript $'$, we see that the offset in the machine words behaves

* In (10) and (10′) we have not made allowances for overflow. This however can be done conventionally by employing saturation where:

$$y(i+1) = 0 \text{ if } \langle \beta y(i) \rangle + m(i) < 0,$$
$$= 2^K - 2^{-F} \text{ if } \langle \beta y(i) \rangle + m(i) > 2^K - 2^{-F},$$

and in every other case (10) holds. Saturation acts to attenuate the offset in the machine words at the two sites.

as follows during epochs of error-free transmission [i.e., periods in which $m(\cdot) = m'(\cdot)$]:

$$|y(i+1) - y'(i+1)| = |f\{y(i)\} - f\{y'(i)\}| \tag{11}$$

[compare with (8)].

The synchronization problem motivates us to impose the following two rather stringent requirements on the behavior of the offset.

Synchronization requirements:

(*i*) The offset is nonincreasing at all instants of error-free transmission.

(*ii*) The integer parts of the machine words at the two sites, and hence the respective log step sizes, differ in at most a finite (preferably small) number of time instants during error-free transmission.

We require the above to hold independent of the statistics of the input process. It is clear from (11) that these requirements imply restrictions on the discrete map f which are investigated below.

Let us digress to better motivate the second of the above requirements. If the integer parts of the machine words at the two sites at any instant are not identical, then the respective log step sizes differ by at least unity and, hence, the ratio of the two step sizes is at least Q [see eq. (3)]; this factor may be unacceptably large since values of Q as high as 1.5 are being

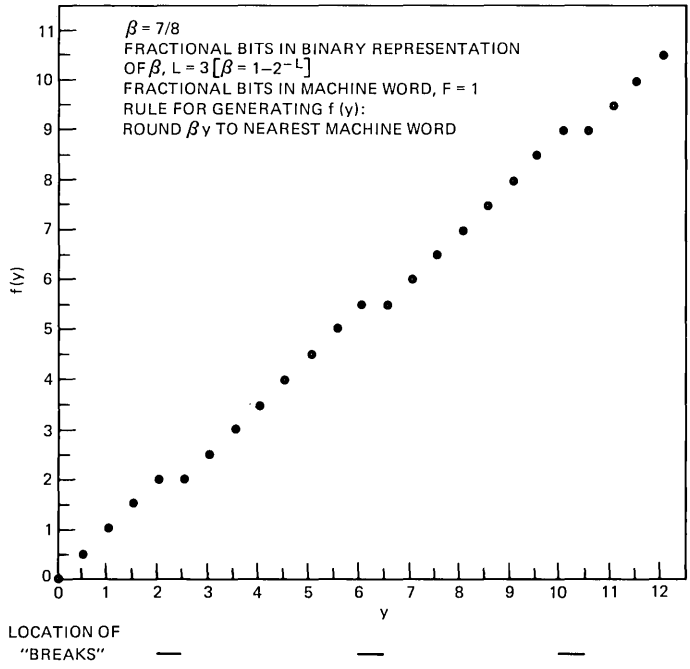


Fig. 2—An example of a naive machine arithmetic.

considered in practical designs.* To illustrate another facet of the second requirement, consider the case where, at a particular instant, the transmitter and receiver machine words are rather close, say, 1.9375 and 2.0625 ($F = 4$). Yet the integer parts are 1 and 2, respectively. Thus the step sizes are Q and Q^2, rather far apart. This example serves to illustrate that the mere proximity of the two machine words is not enough to guarantee that the log step sizes are identical.

In the following discussion, we will need to know the value of L, an integer, which is such that

$$1 - 2^{-L+1} < \beta \leq 1 - 2^{-L};$$

if $\beta = 7/8$, as in Fig. 2, then $L = 3$ and if $\beta = 63/64$ then $L = 6$. To simplify the following discussion, we shall assume that

$$\beta = 1 - 2^{-L}, \tag{12}$$

i.e. $\beta \epsilon \{1/2, 3/4, 7/8, \cdots\}$; with this form for β, L is the minimum number of fractional bits required for the binary representation of β. The assumption on the form of β is unessential, and later in Section 2.2 we indicate that no difficulties are presented if β is not of the assumed form.

We give two different but connected reasons which separately lead to the rather consequential conclusion that $F \geq L$ if the resulting system is to have certain essential properties, including the synchronization capability. The first reason stems directly from the synchronization requirements. We show that the latter requires the map f to incorporate certain contraction properties which in turn can be possible only if the internal machine word has at least L fractional bits. The second related reason is that fewer than L fractional bits gives rise to rounding errors in each iteration of the recursion which makes it hard to predict the effective value of the leakage parameter. Recall from Section 1.1.1 the stringent requirements on the leakage parameter.

Below we amplify both the above arguments. This discussion will motivate a more exact treatment in Section 2.2, which will also provide answers to the questions raised here.

Consider (10′) in conjunction with the synchronization requirements (*i*) and (*ii*). For the first of the synchronization requirements to be satisfied, it is apparent that it is necessary and sufficient that

$$|f(y) - f(y')| \leq |y - y'| \tag{13}$$

for all machine words y and y'. We refer to the above property of the map f as the *weak contraction everywhere* property. The map f shown in Fig.

* This is the case if K is 5 or 6. If K is larger, then it is possible to relax the second requirement by requiring that the offset in the integer parts of the machine words be reduced to a small number (instead of requiring them to be identical). Thus it is possible to trade a higher K for a lower F while keeping $K + F$ fixed. In any case, only minor modifications to the framework that is developed here will allow such cases to be handled.

2 possesses this property by virtue of the fact that the slope of the graph of f is everywhere either 0 (at the breaks) or 1.

For the second of the synchronization requirements to be satisfied, we claim that it is necessary and sufficient that the map f have the following property:

If y and y' are any machine words with different integer parts, then

$$|f(y) - f(y')| \leq \delta |y - y'| \text{ for some } \delta < 1. \tag{14}$$

We call the above the *strong contraction across integer boundaries* property. Sufficiency is clear, since we have that during epochs where the machine words do not have identical integer parts and error-free transmission exists,

$$|y(i) - y'(i)| \leq \delta^i |y(0) - y'(0)|. \tag{15}$$

Conversely, if (14) is not true, then it is easy to construct examples where the integer parts of the two machine words are different at an unbounded number of time instants. Referring to Fig. 2 we see that the graph of f does *not* possess the strong contraction property (14). To illustrate, suppose that initially the two machine words have different integer parts and that both words occur in the range [2.5,6]; we see from the figure that no mechanism exists to prevent the two words from indefinitely remaining in this range and simultaneously having different integer parts.

We will now argue that the above two contraction properties, together with any weak fidelity criterion relating $f(y)$ to βy, implies that $F \geq L$. Observe that the strong contraction property, (14), requires a "break" (see "breaks" in Fig. 2) in the graph of $f(y)$ just prior to *every* integral value of y. Reason: $y = k - 2^{-F}$ and $y = k$, k integral, have different integer parts. Further, if the local slope of the graph of $f(y)$ is not zero, then by virtue of the weak contraction property it is either 1 or -1. Finally, if F fractional bits are used, then each unit interval of y is composed of 2^F intervals of equal length corresponding to that many distinct machine words. These three considerations show that the

$$\text{average slope of the graph of } f(\cdot) \leq \frac{2^F - 1}{2^F} = 1 - 2^{-F}. \tag{16}$$

But $f(y)$ is supposed to approximate βy, $\beta = 1 - 2^{-L}$. Thus, just about any weak fidelity criterion will give that the smallest value of F, which allows the map f to have the properties required of it, is L.

Our second reason is closely related to the aforementioned fidelity criterion. Implicit in a choice of a leakage parameter β with a large number of fractional bits, L, in its binary representation (e.g., $\beta = {}^{63}/_{64}$) is the requirement that the absolute rounding error in each iteration of (10′), $|f\{y(i)\} - \beta y(i)|$, be not larger (at least not by much) than an error in the least significant bit of β, i.e. 2^{-L}:

$$|f(y) - \beta y| < 2^{-L}, \qquad \text{for all machine words } y. \tag{17}$$

Otherwise, there is no *a priori* need to specify β to that degree of precision. (Our experience with the idealized system, discussed previously, shows that it is indeed necessary to specify β to a high degree of precision.) A little thought will convince the reader that for such a bound, (17), on the rounding error to be valid it is necessary that the internal machine word have at least L fractional bits.

In Section 2.2 we show that it is possible to obtain maps f with the weak and strong contraction properties that satisfy the fidelity criterion with the minimum possible number of fractional bits, i.e., $F = L$. We show that, in fact, the maps obtained are *unique*. The results will show that, for our maps, the offset in machine words during error-free transmission decreases exponentially fast to a value less than unity, after which there may be at most $(2^L - 1)$ occasions at which the integer parts differ.

Let us now consider in broad terms what the preceding results imply in terms of the cost and complexity of the digital implementation of the scheme for adaptive quantization discussed in Section 1.1.1. Consider the fairly typical case where the total number of integral log step sizes is 64 and $\beta = {}^{63}/_{64}$, i.e. $K = 6$ and $L = 6$. We now know that the total word length should be at least 12 bits. Consider the implications on the associated ROM size. The table stored in the ROM will have 2^{12} addresses, each address containing 12 bits, giving a total memory size in the transmitter and receiver of about 50K bits each! Moreover, with each additional bit in the internal word, the memory requirement more than doubles.*

In the next section, we propose a new adaptation algorithm and specify the required arithmetic. The new algorithm requires significantly fewer fractional bits in the machine words while possessing the necessary synchronization capability.

II. THE PROPOSED SYSTEM

2.1 Idealized description

We propose the following *interleaved-leakage algorithm* (ILA) as the basis for the machine adaptation of the log step size. For fixed parameters I and γ, $I \geq 2$ and $0 < \gamma < 1$ [see eq. (6)]:

$$\left.\begin{aligned} d(i+1) &= \gamma d(i) + m(i) \\ d(i+2) &= d(i+1) + m(i+1) \\ d(i+I) &= d(i+I-1) + m(i+I-1) \end{aligned}\right\} i = 0, I, 2I, \cdots. \tag{18}$$

* We have considered the possibility of exploiting the idea due to Croisier et al. (Ref. 11) and Peled and Liu (Ref. 12) wherein the ROM size may be reduced at the cost of increased processing time. The processing times available and the relative costs do not make this approach particularly promising at the present time. However, it is an approach worth keeping in mind.

Here γ is the leakage constant, and leakage is introduced only once in every I iterations. Thus we refer to I as the *interleaving interval*. The $m(\cdot)$ terms are the log multipliers, $m(\cdot) \in \{m_1, \cdots, m_N\}$, and the selection rule is as in (6b). However, in general, the optimum values of the multipliers may be different from the ones in the scheme described in Section 1.1.1 (we refer to the latter scheme as the uniform-leakage algorithm, or sometimes only as ULA).

We observe that for two geographically separated implementations, $\{d(\cdot)\}$ and $\{d'(\cdot)\}$, of the recursion in (18) subject to possibly different initial values, $d(0)$ and $d'(0)$, but identical $\{m(\cdot)\}$ sequences, as is the case during error-free transmission, we have for the offset,

$$|d(i) - d'(i)| = (\gamma^{1/I})^i |d(0) - d'(0)|, \qquad i = 0, I, 2I, \cdots. \tag{19}$$

Comparing (19) with the similar expression in (8) for the offset in ULA, we find that the capability for recovery from channel errors is comparable in the two schemes if

$$\gamma^{1/I} = \beta. \tag{20}$$

The above is a key relation. Table I tabulates typical values of β and the corresponding choices of γ and I which give comparable recovery capabilities. There are small, inconsequential errors in the table which has been obtained from the approximation $\gamma = [1 - (1 - \beta)]^I \approx 1 - I(1 - \beta)$ for small values of $(1 - \beta)$.

The important point about the table is that, for given β, the fractional bits required for a binary representation of the equivalent value of γ is reduced by an additional bit for every doubling of the interleaving interval, I, in ILA. This simple fact is at the heart of the system that is proposed.

Table I — Leakage parameters (β,γ) and interleaving intervals (I) for comparable synchronization capabilities in the uniform and interleaved leakage algorithms*

	γ (ILA)				
β (ULA)	$I = 2$	$I = 4$	$I = 8$	$I = 16$	$I = 32$
$127/128$	$63/64$	$31/32$	$15/16$	$7/8$	$3/4$
$63/64$	$31/32$	$15/16$	$7/8$	$3/4$	
$31/32$	$15/16$	$7/8$	$3/4$		

* We have stopped short of using $\gamma = 1/2$ for two reasons. First, there may be no advantage in reducing γ beyond $3/4$ because two fractional bits may be required in any case on account of the specification of the log multipliers, m_r. Second, the change in the step size may be too drastic, and this may be reflected in the subjective quality. However, it is a possibility worth keeping in mind.

A slight generalization of the proposed scheme would have the multiplier set in the iteration where leakage γ is inserted to be different from the common multiplier set in all other iterations. This generalization provides no gain when the midpoint of the input signal intensities ($\hat{\sigma}$ of Section IV) is scaled to be unity, which is the case considered in the simulations reported in Section IV. Goodman[10] has suggested that, when $\hat{\sigma} \pm 1$, the log multipliers in the leaky iterations be $m(\cdot) + (1 - \gamma) \log_Q \hat{\sigma}$, where $\{m(\cdot)\}$ are the log multipliers in the nonleaking iterations.

2.2 The digital implementation

We now consider the digital implementation of the idealized recursion (18).

Here we let L, an integer, be such that $1 - 2^{-L+1} < \gamma \leq 1 - 2^{-L}$. We make the simplifying, and inessential, assumption that $\gamma = 1 - 2^{-L}$; in this case, the binary representation of γ requires L fractional bits. (Later we indicate through an example that it is easy to make the modifications which allow other values of γ to be used.) Assume K integer and L fractional bits for the internal machine words. Thus, following the discussion on the synchronization requirements in Section 1.1.2, we are assuming that the fractional bits in the machine words are the minimum necessary for the system objectives to be satisfied. Finally, assume that the log multipliers $\{m_r\}$ are specified to L fractional bits.

The internal description of the machine is

$$\left.\begin{aligned} y(i+1) &= f\{y(i)\} + m(i) \\ y(i+2) &= y(i+1) + m(i+1) \\ &\vdots \\ y(i+I) &= y(i+I-1) + m(i+I-1) \end{aligned}\right\} \quad i = 0, I, 2I, \cdots, \tag{21}$$

where $y(\cdot)$, the internal machine word, is a $(K + L)$-bit word with L fractional bits. In (21), f maps $(K + L)$-bit words with L fractional bits into other such words. The mapping f may be implemented most easily using ROMs; the characterization of the map f that we give below is a recipe for the programming of the ROMs.*

The integral log step size $d(\cdot)$ is obtained from the internal word $y(\cdot)$ by a rule determined by an *external arithmetic*. We consider two natural and simple external arithmetics, rounding and truncation. Thus,

$$\text{Rounding: } d(\cdot) = [y(\cdot)]_{\text{round}} \tag{22a}$$

$$\text{Truncation: } d(\cdot) = [y(\cdot)]_{\text{truncate}}. \tag{22b}$$

We mean that if, for integral k, $k - 0.5 < y \leq k + 0.5$, then $[y]_{\text{round}} = k$; if $k \leq y < k + 1$ then $[y]_{\text{truncate}} = k$.

* Observe that the specifications of the maps given here and in Appendix A apply as well to the uniform leakage algorithm described in Section 1.1, provided β replaces γ and the appropriate value of the parameter L associated with the leakage parameter β in ULA is substituted.

We consider first the truncating external arithmetic. Following the discussion in Section 1.1.2, we impose the following requirements on the map f. (It is understood that all arguments of the map have L fractional bits.)

$$(i)\quad \forall \sigma_1,\sigma_2, \qquad |f(\sigma_1) - f(\sigma_2)| \leq |\sigma_1 - \sigma_2|:$$

"weak contraction everywhere." (23)

$$(ii)\quad \sigma_1 \epsilon [k, k+1),\ \sigma_2 \epsilon [k+1, k+2),\ k \text{ integral} \Rightarrow \frac{|f(\sigma_1) - f(\sigma_2)|}{|\sigma_1 - \sigma_2|} \leq \delta < 1:$$

"strong contraction across integer boundaries." (24)

$$(iii)\quad \forall \sigma, \qquad |f(\sigma) - \gamma\sigma| < 2^{-L}:$$

"fidelity of discrete map to continuous map." (25)

Recall from Section 1.1.2 that the first two properties are equivalent to the synchronization requirements. We also know that these two conditions together with almost any weak fidelity criterion relating $f(\sigma)$ to $\gamma\sigma$ implies that the number of fractional bits in the machine words is at least L. We find that we can construct maps f which satisfy in addition the fidelity criterion in (*iii*) without incurring the penalty of using more than L fractional bits. Also, as discussed previously, the fidelity criterion in (*iii*) is important in itself.

In Appendix A we give the complete specification of a map for each value of L. In Fig. 3a, we show the graph of the map f for the example of $\gamma = 3/4$, where $L = 2$. In Appendix A we also show that there is only one such map f for any given L which satisfies conditions (*i*) to (*iii*), (23) to (25). Further, for this unique map the value of the contraction parameter δ in (24) is $2\gamma/(1+\gamma)$.

When the external arithmetic is the rounding arithmetic (22a), the

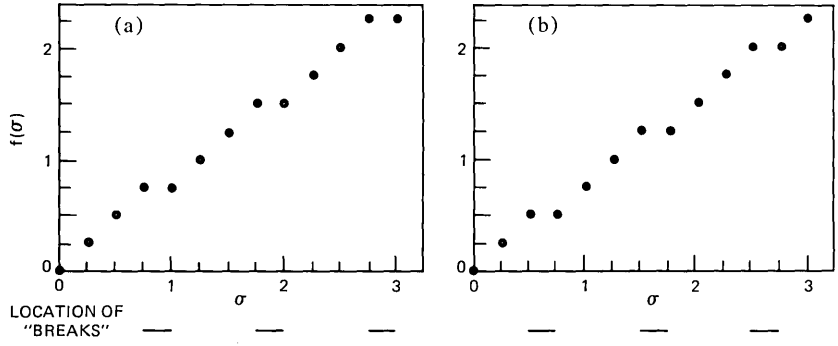


Fig. 3—Machine arithmetics incorporating contraction properties and fidelity criterion for (a) truncating and (b) rounding external arithmetics. $\gamma = 3/4$ and $L = 2$ (see Section 2.2).

resulting map f is somewhat different. Appendix A gives the complete specifications of the maps for all values of L; these maps are also unique. Figure 3b shows the graph of one such map.

Recall that earlier we made the simplifying assumption that $\gamma = 1 - 2^{-L}$. In general, L is defined to be such that $1 - 2^{-L+1} < \gamma \leq 1 - 2^{-L}$. Figure 4 illustrates a map f for the case of $\gamma = 5/8$ ($L = 2$) and the truncating external arithmetic. It may be verified that all the requirements in (23) to (25) are satisfied. We may similarly generate maps satisfying the requirements for arbitrary rational values of γ.

Note that the maps obtained are rather special and quite distinct from the usual maps encountered in digital signal processing.

Another point to note is that while we have specified arithmetics which use the minimum number of fractional bits, $F = L$, additional fractional bits, if they are available, may be put to use by incorporating more than one break in the graph of $f(\sigma)$ per unit interval of σ. The net effect is to give superior synchronization capability.

Finally, note that the implementation of (21) requires by way of hardware only the ROMs, for implementing the map f, and adders. However, the ROMs are used only once in every I iterations. This provides an ideal opportunity for multiplexing the ROMs between different channels and different frequency bands in subband coding[13] applications.

2.3 Synchronization in the digital implementation

We give some bounds on the offset between transmitter and receiver during periods of error-free transmission.

By y and y', two machine words, having different integer parts we mean in the following that $[y]_{\text{round}} \neq [y']_{\text{round}}$ or $[y]_{\text{truncate}} \neq [y']_{\text{truncate}}$, depending on the external arithmetic chosen. Thus, depending upon whether the two machine words have identical or different integer parts, the corresponding log step sizes are identical or different, respectively.

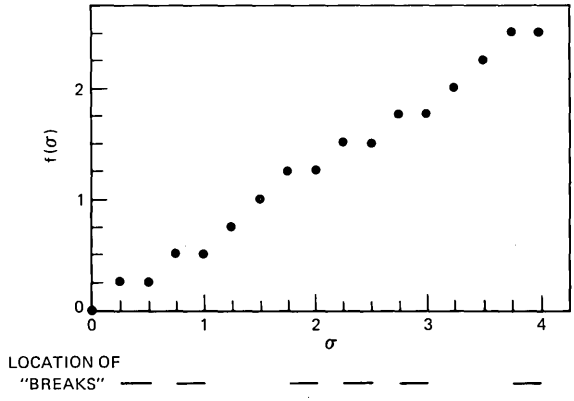


Fig. 4—Machine arithmetic for $\gamma = 5/8$ ($L = 2$) for two fractional bits in machine word and truncating external arithmetic. The contraction requirements and fidelity criterion are satisfied.

Suppose the machine implementations of the recursions in (18) in the transmitter and receiver during error-free transmission are: $i = 0,I,2I,\cdots$

$$y(i+1) = f\{y(i)\} + m(i)$$
$$y(i+2) = y(i+1) + m(i+1)$$
$$y(i+I) = y(i+I-1) + m(i+I-1)$$
$$y'(i+1) = f\{y'(i)\} + m(i)$$
$$y'(i+2) = y'(i+1) + m(i+1)$$
$$y'(i+I) = y'(i+I-1) + m(i+I-1). \quad (26)$$

Observe that

$$|y(i+I) - y'(i+I)| = \cdots = |y(i+1) - y'(i+1)| = |f\{y(i)\} - f\{y'(i)\}|.$$

Now from (23) and (24),

$$|f\{y(i)\} - f\{y'(i)\}|$$
$$\leq |y(i) - y'(i)| \text{ if } y(i) \text{ and } y'(i) \text{ have identical integer parts,} \quad (27)$$
$$\leq \delta|y(i) - y'(i)| \quad \text{if } y(i) \text{ and } y'(i) \text{ have different integer parts.} \quad (28)$$

By repeated application of (28) we see that, if $|y(0) - y'(0)| > 1$, then

$$|y(j) - y'(j)| < 1 \text{ for all } j > I \log\{|y(0) - y'(0)|\}/\log(1/\delta). \quad (29)$$

Thus, once the offset is reduced to less than unity it subsequently remains thus.

Now consider the case where $|y(0) - y'(0)| < 1$. Consider the time instants j which are integral multiples of I. There can be at most $(2^L - 1)$ such time instants at which the integer parts differ. This is because a reduction of 2^{-L} in the offset is guaranteed by (28) in every such time instant. However, at time instants which are not integral multiples of I, the convergence of the integer parts is not quite as strong and is a penalty (which we believe to be insignificant) of ILA.

III. ANALYSIS: PROBABILISTIC ASPECTS

In this section, we investigate the probabilistic behavior of the log step sizes, $\{d(\cdot)\}$, when the input signal variables, $\{x(\cdot)\}$, are random and channel errors are absent. Clearly such an analysis is called for if we are to be able to guarantee certain qualitative features of performance that are basic and necessary in adaptive PCM systems.[4,5] The key notions of the bias function, central log step sizes, and load curves are introduced and their qualitative behavior pinned down.

For our purposes here, the defining equations for the log step sizes are

in (18); the selection rule for the multipliers are in (6b). The key assumption that is made throughout this section is that $\{x(\cdot)\}$ is a sequence of independent, identically distributed random variables with mean zero and standard deviation σ. We sometimes refer to σ as the signal intensity. In keeping with the characteristics of speech, we are interested in σ in the range of $\sigma_{max}/\sigma_{min} = 100$, or even 400 (40 and 52 dB ranges, respectively).

3.1 The bias function

Define the bias function $B(\cdot|\sigma)$ to be

$$B(d|\sigma) \triangleq E[d(i+I)|d(i) = d] - d, \qquad i = 0,I,2I,\cdots. \tag{30}$$

A little thought will show that the right-hand side of (30) does not depend on i—a consequence of the iid assumption on the input signal variables. Different values of σ will generally yield different bias functions, which explains the notation. In engineering parlance, $B(d|\sigma)$ measures, for initial log step size d, the mean drift of the log step size after one cycle of updating of the log step size.

We are able to show for a wide range of values of σ that the bias functions consistently have a distinctive form, depicted in Fig. 5, of considerable significance. In particular, we show that $B(d|\sigma)$ is positive when d is sufficiently small, and negative when d is sufficiently large. Further, under a rather mild restriction, we can prove the consequential result that $B(d|\sigma)$ is monotonic, decreasing with increasing d. The above results in their precise forms are proven in Appendix B. The restriction that is mentioned above is interesting in itself and, roughly, it calls for a propensity for the expected log step sizes after one iteration to be ordered in the same way as the initial log step sizes. This turns out to require, roughly, that $(m_N - m_1)$ be not too large.

The importance of the above results is on account of the following corollary which we state in qualitative terms:

If $(m_N - m_1)$ is not too large, then there exists a unique root, or zero-crossing, of the bias function $B(\cdot|\sigma)$.

Without the monotonicity of the bias function, the possibility exists of

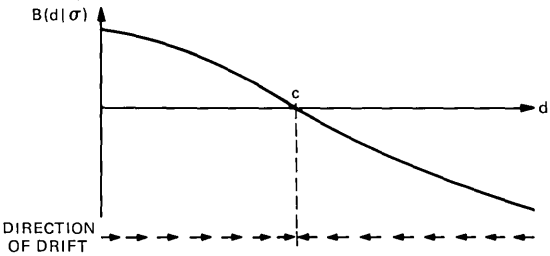


Fig. 5—Sketch of a bias function.

there being many roots with a consequent dilution of the importance that we attach to the root.

Let c denote such a root for a fixed value of σ, Fig. 5:
Definition of c:

$$B(c|\sigma) = 0. \tag{31}$$

We refer to c as the *central log step size* (for signal intensity σ). For a different value of σ and hence a different bias function, the root will generally be different, and to make this dependence quite clear we use the notation $c(\sigma)$.

As the terminology implies, we expect the probability distribution of the log step size to have a concentration of mass around $c(\sigma)$ whenever the signal intensity is σ. The reason for expecting this (see direction of drift indicated by arrows at bottom of Fig. 5) is that, whenever the log step size is not at $c(\sigma)$, the mean drift of the log step size is toward $c(\sigma)$.

The above conclusion is amply borne out by computational results (see Section IV). We find, for instance, that the fit between $c(\sigma)$ and the mean log step size in steady state is extremely good for a rather broad range of values of σ.

In summary, the dual properties of the central log step size (namely, that it predicts so well the mean log step size and that it is so much more tractable and easily obtained) explain the emphasis that we place on the notion of the central log step size.

3.1.1 Method for generating the bias function

The following recursive formula which is developed in Appendix B is the most effective method we know for obtaining the bias function. First, it is necessary to define the following functionals:

$$b_r(\tau) \triangleq 2 \int_{\xi_{r-1}Q^\tau}^{\xi_r Q^\tau} p(\mu)d(\mu), \qquad 1 \le r \le N, \tag{32}$$

where $p(\mu)$ is the common pdf of the input signal variables $\{x(\cdot)\}$. (It is slightly simpler to make as we do the inconsequential assumption that $p(\cdot)$ is symmetrical about 0.) Then $B(d|\sigma)$ is obtained as the solution of the following functional recursion:

$$B_0(d|\sigma) = 0, \ \forall d$$

$$B_k(d|\sigma) = \begin{cases} \sum_{r=1}^{N} b_r(d)\{B_{k-1}(d + m_r|\sigma) + m_r\}, \ 1 \le k \le I - 1 \\ -(1-\gamma)d + \sum_{r=1}^{N} b_r(d)\{B_{k-1}(\gamma d + m_r|\sigma) + m_r\}, \ k = I. \end{cases} \tag{33}$$

Finally, $B(d|\sigma) = B_I(d|\sigma)$.

The above formula is used in the following manner: Assume that the function $B_{k-1}(d|\sigma)$ is known for all values of d. Use (33) to generate next the complete function $B_k(d|\sigma)$. After I such iterations, the resulting function $B_I(d|\sigma)$ is in fact $B(d|\sigma)$.

The reader is referred to eq. (50), Appendix B, for the probabilistic interpretations of the ancillary functions $B_k(\cdot|\sigma)$.

The above formula is used in the analysis presented in Appendix B to determine the previously mentioned qualitative properties of the bias function $B(d|\sigma)$.

Figure 6 is a plot of the bias function $B(d|1)$ for a 16-level quantizer and normally distributed input signal variables. The interleaving interval, I, is 16. Observe in the figure that the graph is for d in the range $[-200,800]$. Values of d outside this range are not of much interest, since the maximum range of the log step sizes in this example is $[\mathrm{Im}_1/(1-\gamma), \mathrm{Im}_N/(1-\gamma)] = [-163,828]$.

3.2 Load curves

The load curves provide information regarding the manner in which the log step sizes depend on the input signal intensity, σ. We use the term to describe a graph of $\log_Q \sigma$ vs. $\bar{d}$, where $\bar{d}$ is the mean log step size in steady state for signal intensity σ. Naturally, the range of σ should cover the range of values expected in the specific application.

From our previous discussion on bias functions and their roots, the

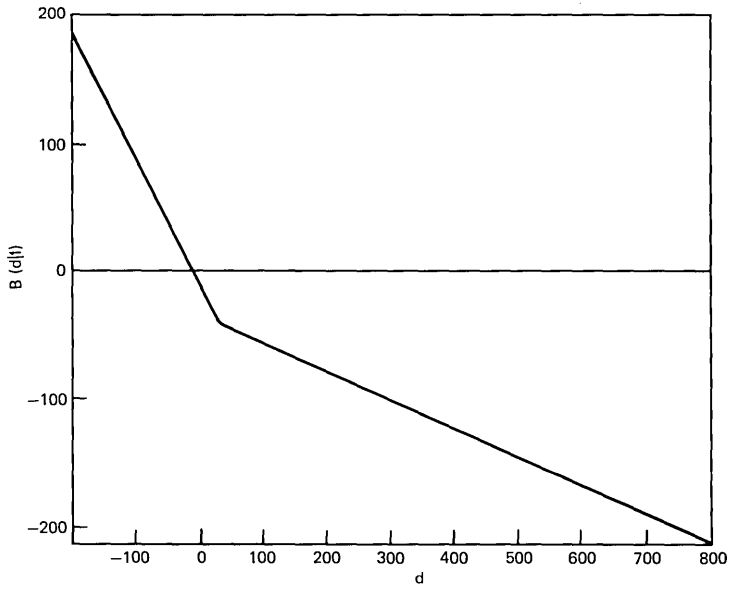


Fig. 6—The bias function for uniform 16-level quantizer and normally distributed input signal variables, $\sigma = 1$. Interleaving internal, $I = 16$ and $\gamma = 0.777$. The log multipliers are given in (39).

central log step sizes, we expect a plot of $\log_Q \sigma$ vs. $c(\sigma)$ to be a rather good fit to the load curves.

The utility of the load curve derives from the fact that it may be visually compared with a plot of the ideal log step size with respect to σ. This information may be obtained from solving a variational problem as is done by Max,[14] who has also tabulated the solutions for the case of normally distributed input signal variables. In any case, the solutions to the variational problem for the *optimum log step size* $\hat{d}(\sigma)$ have the following form

$$\hat{d}(\sigma) = \log_Q \sigma + \hat{D}, \tag{34}$$

where $\hat{D}$ is a constant which depends on the fixed parameters of the quantizer and, importantly, on the common pdf of the input signal variables.

Figure 7 is a plot of the load curve obtained for the 16-level quantizer.

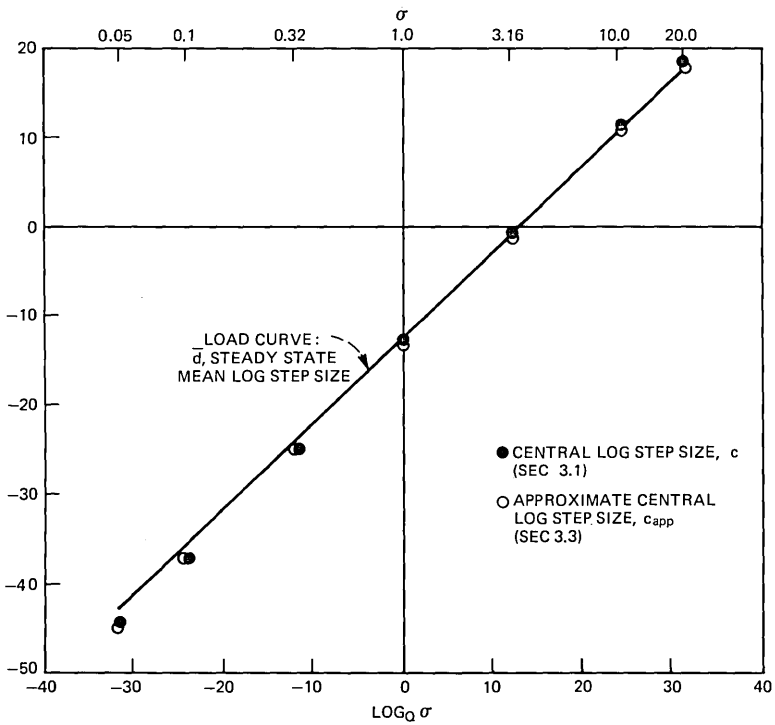


Fig. 7—Load curve ($\bar{d}$), central log step size (c), and approximate log step size (c_{app}) for uniform 16-level quantizer and Gaussian, zero-mean, input signal variables of variance σ^2. The log multipliers are given in (39) and $Q = 1.1$. Interleaving interval, $I = 16$ and leakage, $\gamma = 0.777$.

3.3 The almost-linear dependence of the central log step sizes on signal intensity

Even though a plot of $\log_Q \sigma$ vs $c(\sigma)$ may be expected to be a rather good approximation to, and certainly simpler to obtain than, the load curve ($\log_Q \sigma$ vs $\bar{d}$), it is an unfortunate fact that it is not a very simple matter to obtain $c(\sigma)$. However, our graphs of $c(\sigma)$ have consistently displayed a most remarkable trait, namely, the almost-linearity of $c(\sigma)$ with respect to σ. Intrigued by this feature, we found in an earlier study[9] that it could be explained if the following rather unusual approximation is effective:

$$\int_0^y p(\mu)d\mu \approx \alpha_1 \log y + \alpha_2, \tag{35}$$

where α_1 and α_2 are constants and $p(\cdot)$ is the common pdf of the input signal variables scaled to have unit variance.

Certainly, the above cannot be a good approximation when either y is very small or y is very large. But, as we see in Appendix C, we need the above to be a good approximation only for a limited range of y; specifically, the range of y is required to include the range encountered by $\xi_1 Q^{d(\cdot)}$ at one end, and $\xi_{N-1} Q^{d(\cdot)}$ at the other end, where $d(\cdot)$ is the typical log step size. It turns out that in the important cases where $p(\cdot)$ is either Gaussian or Laplacian, the range of validity of (35) is adequate, at least for the analysis of quantizers with up to 16 levels ($N = 8$). Further details may be found in Ref. 9. For both these distributions, we have found (35) to be an effective approximation in the range $\frac{1}{3} \leq y \leq 2$. For the former distribution, we have found good fits to be obtained if

$$\alpha_1 = 0.44 \text{ and } \alpha_2 = 0.34.$$

(Below, we find it more convenient to express the rhs of (35) as $\alpha_1 \log_Q y + \alpha_2$.)

With (35) as the sole approximation, in Appendix C we go through the involved and tedious process of approximating the bias function and thence deriving its root. The final result, however, is the following remarkably informative formula ($c_{\text{app}}(\sigma)$ is the *approximate central log step size* for signal intensity σ):

$$c_{\text{app}}(\sigma) = S \log_Q \sigma + D, \tag{36}$$

where

$$S = \frac{1}{1 + \dfrac{(1-\gamma)\{1 - 2\alpha_1(m_N - m_1)\}^{I-1}}{1 - \{1 - 2\alpha_1(m_N - m_1)\}^I}} \tag{37}$$

and

$$D = \frac{m_N - 2\sum_{r=1}^{N-1} (m_{r+1} - m_r)(\alpha_1 \bar{\xi}_r + \alpha_2)}{2\alpha_1(m_N - m_1)} S. \tag{38}$$

Let us remark on certain features of the formula. Observe that, on account of α_1 being small, $1 - 2\alpha_1(m_N - m_1) > 0$ almost certainly; for example, $\alpha_1 = 0.018$ when Q [see eq. (3)] is 1.1 and the input signal variables are Gaussian. Consequently, we observe, from the formula in (37) for the slope S, that $S < 1$. Now the ideal slope is 1 [see (34)]. Thus eq. (37) expresses the undesirable but expected fact, alluded to earlier in Section 1.1, that decreasing the leakage parameter γ has the effect of driving the load curve away from the ideal, as sketched in Fig. 8.

As a digression, note that when $\gamma = 1$, the slope S is unity. This is, of course, known to be the case.[4,5] We may also compare the expression for S with a similar expression for ULA derived in Ref. 9—the two expressions are practically identical when $\gamma = \beta^I$ [eq. (20)] and β is close to unity. This important fact, also confirmed in simulations in the example of Section IV, shows that in terms of the loading we expect the behavior in ILA and ULA to be roughly equivalent.

One of the uses that formulas (36) to (38) can be put to is in the optimum choice of the multipliers. The approach we take is that γ and $(m_N - m_1)$ are determined *a priori* on the basis of requirements arising from the quality of synchronization and transient response, respectively. This then fixes the value of S, eq. (37). However, there is still considerable freedom in the choice of the quantities $(m_{r+1} - m_r)$, $1 \leq r \leq N - 1$, and thereby in the choice of the value of D, eq. (38). This degree of freedom may be exploited to determine the point of intersection of the graph of $c_{\text{app}}(\sigma)$ and the ideal graph, which are shown in Fig. 8. A sensible choice for the point of intersection is at the signal intensity, σ, that is most likely to be encountered. Usually,[1] this is at the midpoint of the range of signal intensities expected to be encountered in the application.

IV. COMPUTED RESULTS

Throughout this section, the input signal variables $\{x(\cdot)\}$ are independent, Gaussian, random variables with mean zero and standard deviation σ. The signal intensity σ is varied about a central value of 1.0.

The quantizer is a 16-level, uniform quantizer, i.e., $N = 8$, $\xi_r = r$, $1 \leq r \leq N - 1$, and $\eta_r = r - \frac{1}{2}$, $1 \leq r \leq N$. Throughout, the log base for the step sizes and multipliers, Q, is 1.1.

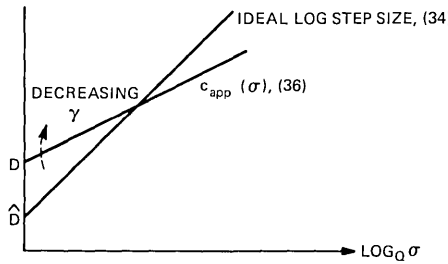


Fig. 8—The behavior of the central log step size compared to the ideal. See eqs. (34) and (36).

For the uniform-leakage algorithm, ULA, we used as the leakage constant $\beta = {}^{63}/_{64}$. The multipliers for ULA are approximately those used by Rosenthal et al.[15] after correction, in the manner suggested in Ref. 1, for the following specifications: In the notation of Ref. 1, $\hat{\sigma}$ = midpoint of signal intensities = 1.0, the ideal loading factor = ideal step size/signal intensity = 0.257. This procedure gave the following values for the log-multipliers for ULA,

$$m(1) = m(2) = m(3) = m(4) = -2.25; \; m(5) = m(6) = 2.50; \; m(7) = 7.25; \; m(8) = 11.50. \quad (39)$$

The multipliers used for the interleaved algorithm, ILA, were also selected to be those given above. We are aware of the advantages of fine tuning the multipliers and Q to take advantage of the special features of ILA, but decided on balance to keep the multipliers and Q unchanged. We found that, as it stands, the transient behavior for ILA is slightly superior to that of ULA; reducing Q in ILA equalizes the transient behavior in the two schemes and yields s/n ratios slightly better than those reported here for ILA.

4.1 Computed load curve, central log step sizes, and their approximation

We illustrate the above notions for the interleaved leakage algorithm for the case of the interleaving interval, $I = 16$. We set $\gamma = \beta^I = 0.777$. Figure 7 plots three quantities with respect to $\log_Q \sigma$: (*i*) $\bar{d}$, the steady-state, mean log step size. This was obtained from 10,000 iterations; (*ii*) $c(\sigma)$, the central log step size defined in (31); (*iii*) $c_{app}(\sigma)$, the approximate central log step size as given by (36) to (38).

For the given specifications,

$$c_{app}(\sigma) = 0.99 \log_Q \sigma - 13.20.$$

To clarify Fig. 7, we have also tabulated in Table II the values of the above variables at seven values of σ.

Table II — Computed load curve, central log step sizes, and their approximation ($I = 16$ and $\gamma = 0.777$)

σ, signal intensity	0.05	0.10	0.3162	1.0	3.162	10.0	20.0
$\bar{d}$, steady state mean log step size	−42.40	−35.53	−24.14	−12.46	−0.84	10.81	17.88
$c(\sigma)$, central log step size	−44.35	−37.07	−25.00	−12.93	−0.85	11.24	18.51
$c_{app}(\sigma)$, approximate central log step size	−44.63	−37.36	−25.28	−13.20	−1.12	10.96	18.23

4.2 S/N ratios and load curve for ULA and ILA

Table III compares signal-to-noise ratios for the two schemes for a variety of interleaving intervals. The signal energy is simply the energy of the variables $\{x(\cdot)\}$. The noise is exactly the difference between the input signal variable and its reconstruction at the receiver, *assuming error-free transmission.* Thus, the reported s/n ratios reflect the effect of the step-size adaptation algorithms but do not measure synchronization capabilities of the systems—the latter is measured separately in Section 4.3.

Note the almost identical s/n ratio performance for the two algorithms, ULA and ILA.

Tables IV and V compare the mean and standard deviations of the log step sizes. Again, note the uniformity of the results for the ULA and ILA; the loading characteristics of the two approaches are almost identical.

Table III — Signal-to-noise ratios (dB)

σ	ULA $\beta = 63/64$	ILA; $I = 2$ $\gamma = \beta^2 =$ 0.969	ILA; $I = 4$ $\gamma = \beta^4 =$ 0.939	ILA; $I = 8$ $\gamma = \beta^8 =$ 0.881	ILA; $I = 16$ $\gamma = \beta^{16} =$ 0.777
0.10	14.89	14.92	14.90	14.70	14.16
0.3162	14.55	14.57	14.56	14.48	14.17
1.0	14.19	14.16	14.18	14.14	14.13
3.162	13.80	13.77	13.63	13.84	13.76
10.0	13.37	13.30	13.36	13.31	13.24

Table IV — Steady-state mean log step sizes

σ	ULA $\beta = 63/64$	ILA; $I = 2$ $\gamma = \beta^2 =$ 0.969	ILA; $I = 4$ $\gamma = \beta^4 =$ 0.939	ILA; $I = 8$ $\gamma = \beta^8 =$ 0.881	ILA; $I = 16$ $\gamma = \beta^{16} =$ 0.777
0.10	−35.75	−35.78	−35.72	−35.66	−35.53
0.3162	−24.12	−24.11	−24.10	−24.13	−24.14
1.0	−12.47	−12.54	−12.47	−12.54	−12.46
3.162	−0.88	−0.90	−0.87	−0.82	−0.84
10.0	10.74	10.81	10.84	10.78	10.81

Table V— Standard deviation of log step size in steady state

σ	ULA $\beta = 63/64$	ILA; $I = 2$ $\gamma = \beta^2 =$ 0.969	ILA; $I = 4$ $\gamma = \beta^4 =$ 0.939	ILA; $I = 8$ $\gamma = \beta^8 =$ 0.881	ILA; $I = 16$ $\gamma = \beta^{16} =$ 0.777
0.10	4.48	4.50	4.57	4.75	5.26
0.3162	4.56	4.63	4.64	4.74	4.97
1.0	4.70	4.69	4.74	4.74	4.85
3.162	4.80	4.80	4.81	4.81	4.80
10.0	4.87	4.90	4.88	4.89	4.97

4.3 The steady-state mean offset in the transmitter and receiver log step sizes

Here we present some computational results connected with the *steady state,* joint distribution of the transmitter and receiver log step sizes assuming, as we have done throughout Section IV, that the input signal variables are independent, normally distributed.

The channel is assumed to be memoryless; further, the event that a transmitted "1" is received as a "0" and the event that a transmitted "0" is received as a "1" have the common probability p. Thus, p is the *bit error probability.* In the numerical results presented below, the following typical value for the bit error probability is assumed: $p = 10^{-4}$.

Two geographically separated implementations of the interleaved leakage algorithm, (18), are assumed to be occurring: $i = 0, I, 2I, \cdots$

$$\begin{aligned} d(i+1) &= \gamma d(i) + m(i) \\ d(i+2) &= d(i+1) + m(i+1) \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ d(i+I) &= d(i+I-1) + m(i+I-1) \end{aligned}$$

$$\begin{aligned} d'(i+1) &= \gamma d'(i) + m'(i) \\ d'(i+2) &= d'(i+1) + m'(i+1) \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ d'(i+I) &= d'(i+I-1) + m'(i+I-1) \end{aligned} \tag{40}$$

The information regarding the log multipliers $m(\cdot)$ are assumed to be coded in the manner shown in Fig. 1 and transmitted through the channel described above. The log multipliers $m'(\cdot)$ are the log multipliers corresponding to the received code word.

By the "steady state mean offset in the transmitter and receiver log step sizes" we mean the quantity $\bar{e}$ where

$$\bar{e} = \lim_{i \to \infty} E\{d(i) - d'(i)\} \tag{41}$$

In Appendix D we show that $\bar{e}$ is given by the following expression:[9]

$$\bar{e} = \frac{I}{1-\gamma} \sum_{r,s=1}^{N} (m_r - m_s) T_{sr} p_r, \tag{42}$$

Table VI — Steady state mean offset in transmitter and receiver log step sizes. Bit error probability in channel, p = 10^{-4}

σ	ULA $\beta = 63/64$	ILA; $I = 2$ $\gamma = \beta^2 =$ 0.969	ILA; $I = 4$ $\gamma = \beta^4 =$ 0.939	ILA; $I = 8$ $\gamma = \beta^8 =$ 0.881	ILA; $I = 16$ $\gamma = \beta^{16} =$ 0.777
0.10	−0.025	−0.025	−0.025	−0.026	−0.026
0.3162	−0.022	−0.022	−0.022	−0.023	−0.024
1.0	−0.020	−0.020	−0.020	−0.021	−0.022
3.162	−0.018	−0.018	−0.018	−0.018	−0.020
10.0	−0.015	−0.015	−0.015	−0.016	−0.017

where $T = \{T_{sr}\}$ is the *channel transition matrix* given below and p_r is the steady state probability that the rth code word is transmitted (00$\cdots$0 is the first code word, 11$\cdots$1 is the last, Nth, code word; the sign bit is ignored).

The channel transition matrix T is defined thus:

$$T_{sr} \triangleq \Pr[s\text{th code word recd.} \mid r\text{th code word trans.}].$$

In the special case where the codes are as shown in Fig. 1, the elements of the matrix are obtained in a simple manner from the Hamming distance between the code words. Thus, if $d(s,r)$ is the Hamming distance between the sth and rth code words, then

$$T_{sr} = p^{d(s,r)}(1-p)^{\log_2 N - d(s,r)}, \qquad 1 \le s,r \le N. \tag{43}$$

In the example under consideration where $N = 8$, $T_{11} = (1-p)^3$, $T_{12} = p(1-p)^2$, etc.

The formula given in (42) for $\bar{e}$, the mean offset in log step sizes, is extremely useful. To see this, recall that $\bar{e}$ is defined in (41) in terms of the joint behavior of the transmitter and receiver in steady state, yet (42) provides the means for calculating $\bar{e}$ provided only that the transmitter log step size distribution is known, since the quantities $\{p_r\}$ are statistics of the latter distribution. Thus, the considerably harder task of evaluating the joint distribution of the log step sizes at the two different sites is circumvented.

Table VI enumerates the computed steady-state mean offset in transmitter and receiver log step sizes for various signal intensities and designs; note the almost identical performance.

V. SUMMARY

We consider it important that digitally implemented adaptive quantization systems possess two properties which, regardless of the statistics of the input signal, ensure that synchronization in the step-size adaptations at the transmitter and receiver is restored during periods of error-free transmission: The offset in step sizes is monotonic and nonincreasing and the step sizes differ in at most a finite number of sampling time instants. A detailed examination of the uniform-leakage algorithm (ULA) shows that a necessary and sufficient condition for the synchronization requirements to be satisfied is that the internal machine arithmetic, given by the nonlinear map f, possesses certain contraction properties. It is further shown that these contraction properties may exist only if the number of fractional bits (F) in the internal machine word is at least L where the leakage parameter β is such that $1 - 2^{-L+1} < \beta \le 1 - 2^{-L}$. Thus, if $\beta = 1 - 2^{-L}$ then L is the number of fractional bits required for the binary representation of β. We proceed to show that it is actually possible to obtain internal machine arithmetics which satisfy

all the requirements with the minimum possible number of fractional bits, i.e., $F = L$. The arithmetics that we obtain are moreover unique. With these arithmetics the offset in machine words during error-free transmission decreases exponentially fast to a value less than unity, after which there may be at most $(2^L - 1)$ occasions in which the step sizes differ.

We give a complete specification of the unique maps f. Thus, in the case where truncation is used to obtain the log step size from the internal machine word, the formula that generates f is:

If $\sigma = k + j2^{-L}$, where k and j are integral and $0 \leq j \leq 2^L - 1$, then

$$f(\sigma) = k(1 - 2^{-L}) + j2^{-L}.$$

Figure 3a is the graph of the map f for the example of $L = 2$.

Even the minimum length of the machine words translate into large memory requirements in ROM-based implementations. Thus, in the fairly typical case where the total number of step sizes is 64 and the leakage parameter $\beta = {}^{63}/_{64}$, we find that the minimum word length is 12 bits, which translates into a ROM size of about 50K bits.

We propose a new adaptation algorithm which is considerably more efficient in terms of the memory used in the implementation. In this algorithm, ILA, leakage is interleaved infrequently but at regular intervals into the recursion for the step-size adaptation. Thus, this scheme has as parameters γ, the leakage parameter, and I, the interleaving interval. We find that, for comparable synchronization capabilities in ULA and ILA, the parameters are related thus:

$$\gamma^{1/I} = \beta.$$

Thus for β close to unity, $\gamma \approx 1 - I(1 - \beta)$. Table I shows that for given β the fractional bits required for the binary representation of the equivalent value of γ is reduced by an additional bit for every doubling of the interleaving interval.

To illustrate, consider the example given above where $\beta = {}^{63}/_{64}$; the new scheme provides the option of interleaving leakage once in 8 iterations ($I = 8$) with a leakage parameter $\gamma \approx {}^{7}/_{8}$, which has three fractional bits. Thus, for the same total number of step sizes, the total word length required is 9 bits, which translates into an ROM size of about 5K bits and an order-of-magnitude reduction in memory size. Furthermore, the essential costly element of the system, the ROM, is used only once in 8 iterations, thus allowing for the additional multiplexing of the ROM.

The internal machine arithmetic that is proposed for ILA is identical to that specified for ULA, except that the machine word in the former system is of shorter length.

A detailed theoretical analysis of the statistical behavior of the step sizes for independent random inputs is undertaken. Perhaps the most

insightful result obtained is a simple formula giving the approximate dependence on the input signal intensity, σ, of the central log step size, $c(\sigma)$, which is the particular log step size about which the distribution of log step sizes is concentrated. The formula depends on only two parameters, α_1 and α_2, of the input signal distribution; in the case of Gaussian input distributions, $\alpha_1 \approx 0.44 \log Q$ and $\alpha_2 \approx 0.34$. This simple formula is given in (36) to (38).

The idealized adaptation algorithms were simulated for a representative 16-level quantizer and independent, Gaussian inputs. In the simulations, the multipliers in ILA were selected to be identical to those used in ULA, although *in general we expect the optimal multipliers to be different for the two schemes.* The results of the simulations show that the performances of the systems are almost identical.

APPENDIX A

Specification of the Machine Arithmetics

We describe first the maps f corresponding to the truncating external arithmetic in (22b) which satisfy conditions (*i*) to (*iii*) given in (23) to (25), Section 2.2. In the example shown in Fig. 3a, observe that the breaks, i.e., zero slope segments between pairs of points, occur just prior to the integral values of σ. This is also the rule by which f is obtained for general values of L.

The following formula generates f for general values of L:

$$\text{If } \sigma = k + j2^{-L}, k \text{ and } j \text{ integral and } 0 \le j \le 2^L - 1, \tag{44}$$

then $f(\sigma) = k(1 - 2^{-L}) + j2^{-L}$.

Condition (*i*), (23), is trivially verified. For condition (*ii*), (24), note that for all integral k

$$f(k + 1 - 2^{-L}) - f(k + 1) = 0. \tag{45}$$

Thus a strong contraction across integer boundaries exists and, in fact, for σ_1 and σ_2 with different integer parts

$$\frac{|f(\sigma_1) - f(\sigma_2)|}{|\sigma_1 - \sigma_2|} \le \frac{2\gamma}{1 + \gamma}, \tag{46}$$

so that we may take

$$\delta = 2\gamma/(1 + \gamma) < 1. \tag{47}$$

For the final condition (*iii*), we find that

$$0 \le f(\sigma) - \gamma\sigma \le 2^{-L}(1 - 2^{-L}), \tag{48}$$

where the two inequalities become equalities at $\sigma = k$ and $\sigma = k - 2^{-L}$, respectively, whenever k is integral.

We can also show rather easily that the map f given by (44) is unique,

i.e., there does not exist any other map satisfying the requirements (*i*) to (*iii*). Uniqueness follows from the following two reasons: (*a*) Condition (*ii*) requires that there be a break in the graph of f between $\sigma = k - 2^{-L}$ and $\sigma = k$, k integral, i.e., $f(k - 2^{-L}) = f(k)$. Reason: $\sigma = k - 2^{-L}$ and $\sigma = k$ have different integer parts. (*b*) In order to satisfy at once both the fidelity condition (*iii*) and the weak contraction (*i*) there can be at most one break in the typical integer interval $[k, k + 1]$.

We now describe the slightly different map f which is obtained for the rounding external arithmetic, (22a). For the requirements on f, the only difference is in condition (*ii*) which now reads as follows:

$$(ii')\ \begin{matrix}\sigma_1 \epsilon (k - 1/2, k + 1/2] \\ \sigma_2 \epsilon (k + 1/2, k + 3/2]\end{matrix} \Rightarrow \frac{|f(\sigma_1) - f(\sigma_2)|}{|\sigma_1 - \sigma_2|} \leq \delta < 1. \qquad (24')$$

The graph of f shown in Fig. 3b is obviously similar to the one displayed in Fig. 3a, the main difference being the locations of the breaks which are here positioned immediately following the midpoint of the integer intervals.

We rapidly summarize the key features of f. The formula for generating f for general L is:

If $\sigma = k - 1/2 + j2^{-L}$, $\quad k$ and j integral, $1 \leq j \leq 2^L$,

then $$f(\sigma) = k(1 - 2^{-L}) - 1/2 + j2^{-L}. \qquad (44')$$

The weak contraction condition (*i*) is trivially satisfied as well as the strong contraction condition (*ii'*), (24'), with the same value of δ that was previously obtained:

$$\delta = 2\gamma/(1 + \gamma) < 1. \qquad (47')$$

Finally,

$$|f(\sigma) - \gamma\sigma| \leq 2^{-L-1}, \qquad (48')$$

and hence condition (*iii*) is also satisfied. It is noteworthy that in keeping with the familiar properties of rounding and truncating, the above error bound is generally smaller than the corresponding bound in (48) for the truncating external arithmetic.

The arguments used previously for establishing uniqueness apply as well for the above construction.

APPENDIX B

On the Bias Function

We give here the derivations of the results on the bias function that are stated in Section 3.1, accompanied by more detailed insights and interpretations. It is convenient to drop the adjunct σ in $B(\cdot|\sigma)$, the bias function, with the understanding that here σ is arbitrary, but fixed.

B.1 Generating the bias function

We derive (33), which is a functional recursion yielding the bias function,

$$B(d) = E[d(i)|d(0) = d] - d. \tag{49}$$

Define the ancillary functions

$$B_k(d) \triangleq E[d(I)|d(I-k) = d] - d,\ 0 \le k \le I, \tag{50}$$

so that

$$B(d) = B_I(d).$$

Observe that

$$\begin{aligned} E[d(I)|d(I-k) = d] &= \sum_s s\Pr[d(I) = s|d(I-k) = d] \\ &= \sum_t \Pr[d(I-k+1) = t|d(I-k) = d] \\ &\quad \times E[d(I)|d(I-k+1) = t], \end{aligned} \tag{51}$$

where the Markov property has been used to obtain (51). Now t can take only N possible values. In fact, from (18), we see that if $k < I$, then $t \epsilon \{d + m_r | r = 1, \cdots, N\}$, and if $k = I$ then $t \epsilon \{\gamma d + m_r | r = 1, \cdots, N\}$. Further, the respective probabilities are easily given in terms of the functionals $b_r(y)$, $1 \le r \le N$, defined in (32), of the common pdf of the input signal variables. Thus,

$$\begin{aligned} b_r(d) &= \Pr[\xi_{r-1}Q^d \le |x(\cdot)| < \xi_r Q^d] \\ &= \begin{cases} \Pr[d(i-k+1) = d + m_r | d(I-k) = d], \\ 1 \le k \le I-1 \\ \Pr[d(I-k+1) = \gamma d + m_r | d(I-k) = d], \\ k = I. \end{cases} \end{aligned} \tag{52}$$

Substituting in (51), we arrive at the relations

$$\begin{aligned} &E[d(I)|d(I-k) = d] \\ &\quad = \begin{cases} \sum_{r=1}^{N} b_r(d) E[d(I)|d(I-k+1) = d + m_r], & 1 \le k \le I-1 \\ \sum_{r=1}^{N} b_r(d) E[d(I)|d(I-k+1) = \gamma d + m_r], & k = I. \end{cases} \end{aligned} \tag{53}$$

Substituting in the expressions in (50) for the functions $B_k(\cdot)$, we obtain the recursive formula given in the main text:

$$B_0(d) \equiv 0,$$

$$B_k(d) = \begin{cases} \sum_{r=1}^{N} b_r(d)\{B_{k-1}(d+m_r)+m_r\}, & 1 \le k \le I-1 \quad (54) \\ -(1-\gamma)d + \sum_{r=1}^{N} b_r(d)\{B_{k-1}(\gamma d+m_r)+m_r\}, & k = I, \quad (55) \end{cases}$$

and $B(d) \equiv B_I(d)$.

B.2 The range of the bias function

Note that, as $d \to -\infty$, the values of all the probabilities $b_1(d), \cdots, b_{N-1}(d)$ approach 0, while $b_N(d) \to 1$. Similarly, as $d \to \infty$, the values of all the probabilities $b_2(d), \cdots, b_N(d)$ approach 0, while $b_1(d) \to 1$. Thus, from (54) we have that

$$\text{As } d \to -\infty, B_1(d) \to m_N, \qquad \text{and as } d \to \infty, B_1(d) \to m_1. \quad (56)$$

Iterating, we obtain that

$$\text{As } d \to -\infty, B_{I-1}(d) \to (I-1)m_N, \quad \text{as } d \to \infty, B_{I-1}(d) \to (I-1)m_1. \quad (57)$$

Finally, for the bias function we obtain from the above and (55) that

$$\begin{aligned} d \to -\infty, \qquad & B(d) \approx -(1-\gamma)d + \mathrm{Im}_1 > 0 \\ d \to \infty, \qquad & \approx -(1-\gamma)d + \mathrm{Im}_N < 0. \end{aligned} \quad (58)$$

The above is the basis for the claim that at least one zero-crossing of the bias function is guaranteed from observing the values of the function at the two limits.

B.3 The monotonicity of the bias function

We establish here sufficient conditions which imply the rather important monotonicity property of the bias function. Equations (54) and (55) provide the working definition of the bias function. Observe from (54) that for $1 \le k \le I-1$,

$$\begin{aligned} B_k'(d) &= \sum_{r=1}^{N} b_r'(d)\{B_{k-1}(d+m_r)+m_r\} + \sum_{r=1}^{N} b_r(d)B_{k-1}'(d+m_r) \\ &= -\sum_{r=1}^{N-1} F_r'(d)\{B_{k-1}(d+m_{r+1}) - B_{k-1}(d+m_r) + m_{r+1} - m_r\} + \prime\prime. \end{aligned} \quad (59)$$

We have found it convenient to introduce

$$F_r(d) \triangleq \sum_{s=1}^{r} b_s(d), \qquad 1 \le r \le N. \quad (60)$$

The reason for this is that $F_r'(d)$ is positive since

$$F_r(d) = 2\int_0^{\xi_r Q^d} p(\mu)d\mu. \quad (61)$$

At this point, it is worth noting from (59) that $B_{k-1}' < 0$ is not enough

to establish that $B_k' < 0$; it is necessary in addition that B_{k-1}' be not excessively negative. This motivates the bounding of the derivative of B_{k-1} from both below and above. We therefore introduce the quantities

$$\alpha_k \leq \min_y B_k'(y); \quad \max_y B_k'(y) \leq \beta_k, \tag{62}$$

where it is understood that we are only interested in y having values in the finite dynamic range of the log step size. Further, let

$$\delta(d) \triangleq \sum_{r=1}^{N-1} F_r'(d)(m_{r+1} - m_r) \tag{63}$$

and

$$0 < \delta_{\min} \leq \delta(d) \leq \delta_{\max}. \tag{64}$$

From (59) we obtain

$$B_k'(d) \leq -\delta(d)(\alpha_{k-1} + 1) + \beta_{k-1}$$

$$\leq -\delta_{\min}(\alpha_{k-1} + 1) + \beta_{k-1}, \textit{ assuming } \alpha_{k-1} \leq -1.$$

Thus, we may take

$$\beta_k = -\delta_{\min}(\alpha_{k-1} + 1) + \beta_{k-1}, \tag{65}$$

provided $\alpha_{k-1} \leq -1$. In identical fashion, we also obtain

$$\alpha_k = -\delta_{\max}(\beta_{k-1} + 1) + \alpha_{k-1}, \tag{66}$$

again assuming $\alpha_{k-1} \leq -1$.

Summarizing, we have at this stage a coupled pair of recursions for the upper and lower bounds on the derivatives of the functions B_k, $1 \leq k \leq I-1$, provided $\alpha_{k-1} \geq -1$, $1 \leq k \leq I-1$. Finally, we also have from (55) that

$$B'(d) = B_I'(d) \leq (1 - \gamma) - \delta_{\min}(\alpha_{I-1} + 1) + \gamma\beta_{I-1}. \tag{67}$$

We may now solve the linear recursions in (65) and (66) for (α_k, β_k) with the initial conditions $\alpha_0 = \beta_0 = 0$. The following solution is obtained: $1 \leq k \leq 1$,

$$\alpha_k = \frac{1}{2}\{(1 + \bar{\delta})^k + (1 - \bar{\delta})^k\} - \frac{1}{2} \cdot \frac{\delta_{\max}}{\bar{\delta}} \cdot \{(1 + \bar{\delta})^k - (1 - \bar{\delta})^k\} - 1. \tag{68}$$

$$\beta_k = \frac{1}{2}\{(1 + \bar{\delta})^k + (1 - \bar{\delta})^k\} - \frac{1}{2} \cdot \frac{\bar{\delta}}{\delta_{\max}} \cdot \{(1 + \bar{\delta})^k - (1 - \bar{\delta})^k\} - 1. \tag{69}$$

We have denoted by $\bar{\delta}$ the geometric mean of $\delta_{\max}$ and $\delta_{\min}$, i.e.,

$$\bar{\delta} = (\delta_{\min}\delta_{\max})^{1/2}. \tag{70}$$

The reader will recall that the recursions (65) and (66) were contingent

upon $\alpha_{k-1} \geq -1$. We find, upon examining the "solutions," that we can ensure its validity over the range $1 \leq k \leq I-1$ provided $\alpha_{I-1} \geq -1$, i.e.,

$$\delta_{\max} \leq \bar{\delta} \cdot \frac{(1+\bar{\delta})^{I-1} + (1-\bar{\delta})^{I-1}}{(1+\bar{\delta})^{I-1} - (1-\bar{\delta})^{I-1}}. \tag{71}$$

The above is a key relation. The first observation on it is that the relation implies not only that $\alpha_{I-1} \geq -1$ but also that $\beta_{I-1} \leq 0$, which is of primary interest. This may be verified either directly from the expression in (69) or, more conveniently, from the recursion in (65) for β_k and the fact that $\beta_0 = 0$. But, as an examination for the bound on $B'(d)$ in (67) shows, these two conclusions, namely, $\alpha_{I-1} \geq -1$ and $\beta_{I-1} \leq 0$, are sufficient to guarantee that $B'(d) < 0$. We have thus arrived at the main result of this section:

$$\text{If } \delta_{\max} \text{ satisfies the inequality (71), then } B'(d) < 0. \tag{72}$$

Some insight into the nature of the inequality (71) may be gained by considering the case of $\bar{\delta} \ll 1$. In this case, the rhs of (71) reduces to $1/(I-1)$. Further, we observe from (68) and (69) that $\alpha_k \approx -k\delta_{\max}$ and $\beta_k \approx -k\delta_{\min}$. Thus, summarizing, we have that

$$\text{If } \bar{\delta} \ll 1 \text{ then } \alpha_k \approx -k\delta_{\max}, \quad \beta_k \approx -k\delta_{\min}, \quad 1 \leq k \leq I-1 \text{ and (71) requires that } \delta_{\max} \leq 1/(I-1). \tag{73}$$

Thus, we have demonstrated that the monotonicity of the bias function is implied if the quantity $\delta(d)$ defined in (63) is uniformly small.

Let us now examine the probabilistic import of the condition in (71), namely, that

$$\delta(d) = \Sigma F_r'(d)(m_{r+1} - m_r)$$

be not large. First, recall from the definition of $F_r(d)$ in (61) that

$$F_r'(d) = 2(\ln Q)(\xi_r Q^d)p(\xi_r Q^d), \qquad 1 \leq r \leq N-1. \tag{74}$$

Thus,

$$\begin{aligned}\delta(d) &= 2(\ln Q) \sum_{r=1}^{N-1} (m_{r+1} - m_r)(\xi_r Q^d)p(\xi_r Q^d) \\ &= 2 \sum_{r=1}^{N-1} \ln (M_{r+1}/M_r)(\xi_r Q^d)p(\xi_r Q^d).\end{aligned} \tag{75}$$

Requiring that $\delta(d)$ be not too large is tantamount to requiring that the ratios of the multipliers, M_{r+1}/M_r, be not too large. To make this connection quite transparent, we see that

$$\delta(d) \leq 2 \ln (M_N/M_1) \left[\max_y yp(y) \right]. \tag{76}$$

For $p(\cdot)$ Gaussian with variance σ^2, observe that

$$\max_y yp(y) = p(\sigma) = 0.242, \tag{77}$$

so that, in this case, (76) states that

$$\delta(d) \leq 0.484 \ln (M_N/M_1). \tag{78}$$

The above is not a particularly good bound, relative to the expression in (75), but it does illuminate the manner in which $\delta_{\max}$ depends on the ratios of the multipliers.

Finally, in summary let us recall in purely qualitative terms the reasons for requiring that $\delta(d) = \Sigma F_r'(d)(m_{r+1} - m_r)$ be not large. This condition is tied in a natural way to the conditions that $B_k'(d) \geq -1$, $1 \leq k \leq I - 1$, which is at the core of the above analysis since it follows rather easily from these conditions that $B_k'(d) \leq 0$, also. The conditions "$B_k'(y) \geq -1$" have an entirely natural, underlying probabilistic interpretation. It merely states that, for two starting log step sizes, $d(0) = d$ and $\dot{d}(0) = d'$, where, say, the ordering is $d < d'$, the respective expected log step sizes after k iterations should also be ordered in the same way. A little thought is enough to convince one that such a condition can only be guaranteed by requiring that $\delta(d)$ be not too large, since $\delta(d)$ itself measures the potential for initial orderings to be reversed in one iteration.

APPENDIX C

Approximate Formula for the Central Log Step Sizes

The object here is to derive the following approximate formula for the dependence of the central log step size on the signal intensity, σ:

$$c_{\text{app}}(\sigma) = S \log_Q \sigma + D, \tag{79}$$

where S and D, given in (37) and (38), are obtained from the fixed parameters of the system. The sole approximation that is made is in approximating the distribution of the input signal variables in the following manner:

$$\int_0^y p(\mu)d\mu \approx \alpha_1 \log_Q y + \alpha_2, \tag{80}$$

where $p(\cdot)$ is the pdf of the input signal variables normalized to have unit variance.

The procedure that is followed consists of first deriving the approximation to the bias function, using the recursive formula in (33), and subsequently deriving the root of the approximate bias function. Observe that the recursive formula in (33) calls for the quantities $b_r(\cdot)$, $1 \leq r \leq N$. We find it essential to work with the partial sums

$$F_r(d) = \sum_{s=1}^{r} b_s(d), \qquad 1 \le r \le N-1$$
$$= 2\int_0^{\xi_r Q^d} p(\mu)d\mu$$
$$\approx 2\alpha_1 \log_Q(\xi_r Q^d/\sigma) + 2\alpha_2, \text{ from (80)},$$
$$= 2\alpha_1 d - 2\alpha_1 \log_Q \sigma + (2\alpha_2 + 2\alpha_1 \bar{\xi}_r), \tag{81}$$

where σ^2 is the variance of the input signal variables. Note that $F_N(d) = 1$.

Examining (33), we find that we may also write it as follows [for notational simplicity, we drop the adjunct σ in $B_k(d \mid \sigma)$]:
for $1 \le k \le I-1$

$$B_k(d) = B_{k-1}(d + m_N) + m_N - \sum_{r=1}^{N-1} F_r(d)\{B_{k-1}(d + m_{r+1}) - B_{k-1}(d + m_r) + m_{r+1} - m_r\}. \tag{82}$$

Now suppose that $B_{k-1}(d)$ may be expressed in the form

$$B_{k-1}(d) = (f_{k-1} - 1)d + g_{k-1} \log_Q \sigma + h_{k-1}, \tag{83}$$

where $(f_{k-1}, g_{k-1}, h_{k-1})$ do not depend on either d or σ. Certainly, $B_0(d)$ may be expressed in this form since $B_0(d) \equiv 0$. We now show that $B_k(d)$ may also be expressed in the above manner.

Upon substituting the above expression for $B_{k-1}(d)$ and the expression in (81) for $F_r(d)$, in (82) we find that

$$B_k(d) = (f_k - 1)d + g_k \log_Q \sigma + h_k, \tag{84}$$

where

$$f_k = \{1 - 2\alpha_1(m_N - m_1)\}f_{k-1},$$
$$g_k = g_{k-1} + 2\alpha_1(m_N - m_1)f_{k-1},$$
$$h_k = h_{k-1} + \left\{m_N - 2\sum_{1}^{N-1}(m_{r+1} - m_r)(\alpha_1\bar{\xi}_r + \alpha_2)\right\}f_{k-1}. \tag{85}$$

Certainly, the newly defined quantities are independent of d and $\log_Q \sigma$. Thus, the basis exists for an inductive construction. Further, the coupled recursions in (85) are trivial to solve for the initial conditions $f_0 = 1, g_0 = 0, h_0 = 0$; thus, we obtain $(f_{I-1}, g_{I-1}, h_{I-1})$.

As is apparent from (23), the final iteration in the recursion for generating the bias function differs from all the others. In fact,

$$f_I = \{\gamma - 2\alpha_1(m_N - m_1)\}f_{I-1}$$
$$g_I = g_{I-1} + 2\alpha_1(m_N - m_1)f_{I-1}$$
$$h_I = h_{I-1} + \left\{m_N - 2\sum_{1}^{N-1}(m_{r+1} - m_r)(\alpha_1\bar{\xi}_r + \alpha_2)\right\}f_{I-1}. \tag{86}$$

The complete solution for the approximation to the bias function is:

$$B(d) = (f_I - 1)d + g_I \log_Q \sigma + h_I, \tag{87}$$

where

$$f_I = -(1-\gamma)\{1 - 2\alpha_1(m_N - m_1)\}^{I-1} + \{1 - 2\alpha_1(m_N - m_1)\}^I, \tag{88}$$

$$g_I = 1 - \{1 - 2\alpha_1(m_N - m_1)\}^I,$$

$$h_I = \frac{\left\{m_N - 2\sum_{1}^{N-1}(m_{r+1} - m_r)(\alpha_1 \bar{\xi}_r + \alpha_2)\right\}\{1 - \{1 - 2\alpha_1(m_N - m_1)\}^I\}}{2\alpha_1(m_N - m_1)}$$

Recall that the central log step size is the root of the bias function $B(d)$. Thus, denoting by $c_{\text{app}}(\sigma)$ the root of the function in (87), we obtain

$$c_{\text{app}}(\sigma) = \frac{g_I}{1 - f_I} \log_Q \sigma + \frac{h_I}{1 - f_I} \tag{89}$$

$$= S \log_Q \sigma + D, \tag{90}$$

where S and D, trivially identified by comparing the two expressions, are as given in the main text, (37) and (38).

APPENDIX D

Formula for the Steady State, Mean Offset in Transmitter and Receiver Log Step Sizes

We derive the formula for $\bar{e}$ given in (42). First, it is necessary to define certain quantities in connection with (40), which describes the step-size adaptations at the two sites.

$$e(\cdot) \triangleq d(\cdot) - d'(\cdot), \quad \text{the offset at time } \cdot, \tag{91}$$

and $u(\cdot) \triangleq m(\cdot) - m'(\cdot)$, the offset in the log multipliers at time $\cdot$. From (40) we obtain

$$\left.\begin{aligned} e(i+1) &= \gamma e(i) + u(i) \\ e(i+2) &= e(i+1) + u(i+1) \\ &\cdots\cdots\cdots\cdots \\ e(i+I) &= e(i-I-1) + u(i+I-1) \end{aligned}\right\} \qquad i = 0, I, 2I, \cdots. \tag{92}$$

Thus,

$$e(i+I) = \gamma e(i) + \{u(i) + u(i+1) + \cdots + u(i+I-1)\}. \tag{93}$$

Taking expectations of both sides of the equation,

$$\bar{e}(i+I) = \gamma \bar{e}(i) + \{\bar{u}(i) + \bar{u}(i+1) + \cdots + \bar{u}(i+I-1)\}, \tag{94}$$

where the bar has been used to denote mean values.

Consider $\bar{u}(i)$, the first term inside the parentheses. Observe that $u(\cdot) \in \{m_r - m_s \mid 1 \le r,s \le N\}$. Also,

$$\bar{u}(i) = \sum_{r,s=1}^{N} (m_r - m_s) \Pr\left[\begin{array}{l} r\text{th code word transmitted } \textit{and } s\text{th} \\ \text{code word received at time } i. \end{array}\right]$$

$$= \sum_{r,s=1}^{N} (m_r - m_s) \Pr\left[\begin{matrix} s\text{th code word recd.} \\ \end{matrix} \middle| \begin{matrix} r\text{th code} \\ \text{word trans.} \end{matrix}\right]$$

$$\times \Pr\left[\begin{matrix} r\text{th code word trans.} \\ \text{at time } i \end{matrix}\right] = \sum_{r,s=1}^{N} (m_r - m_s) T_{sr} p_r(i), \quad (95)$$

where T_{sr} is simply the (s,r)th element of the channel transition matrix, and $p_r(i)$, $1 \le r \le N$, is simply obtained from the pdf of the transmitter log step size at time i.

Expressions for $\bar{u}(i+1), \cdots, \bar{u}(i+I-1)$ may similarly be derived. Thus, for $i = 0, I, 2I, \cdots$

$$\bar{u}(i) + \cdots \bar{u}(i+I-1) = \sum_{r,s=1}^{N} (m_r - m_s) T_{sr} \{p_r(i) + \cdots + p_r(i+I-1)\}. \quad (96)$$

To proceed further, it is necessary to assume ergodicity, i.e., more specifically, convergence in the mean for the time-evolving distributions of the transmitter log step size. With this assumption, as $i \to \infty$

$$\bar{e}(i) \to \bar{e} \quad (97)$$

and

$$p_r(i) + \cdots + p_r(i+I-1) \to I p_r, \qquad 1 \le r \le N, \quad (98)$$

where $\bar{e}$ and p_r have the interpretations mentioned in the main text. Substituting in (94) and (96) yields

$$\bar{e} = \frac{I}{1-\gamma} \sum_{r,s=1}^{N} (m_r - m_s) T_{sr} p_r, \quad (32)$$

which is what we set out to establish.

REFERENCES

1. D. J. Goodman and R. M. Wilkinson, "A Robust Adaptive Quantizer," IEEE Trans. Communication (Correspondence) (November 1975), pp. 1362–1365.
2. N. S. Jayant, "Adaptive Quantization with a One-Word Memory," B.S.T.J., *52*, No. 7 (September 1973), pp. 1119–1144.
3. R. M. Wilkinson, "An Adaptive Pulse Code Modulator for Speech," Proc. Int. Conf. Commun., Paper 1C (June 1971), pp. 1.11–1.15.
4. D. Mitra, "New Results From A Mathematical Study of an Adaptive Quantizer," B.S.T.J., *54*, No. 2 (February 1975), pp. 335–368.
5. D. J. Goodman and A. Gersho, "Theory of an Adaptive Quantizer," IEEE Trans. Commun., *COM-22* (August 1974), pp. 1037–1045.
6. P. Cumminskey, N. S. Jayant, and J. L. Flanagan, "Adaptive Quantization in Differential PCM Coding of Speech," B.S.T.J., *52*, No. 7 (September 1973), pp. 1105–1118.
7. S. Bates, unpublished work.
8. R. Steele, *Delta Modulation Systems*, New York: John Wiley, 1975.
9. D. Mitra, "An Almost Linear Relationship Between the Step Size Behavior and the Input Signal Intensity in Robust Adaptive Quantization," to be presented at the National Telecommunications Conference, 1978.
10. D. J. Goodman, private communication.

11. A. Croisier, D. J. Esteban, M. E. Levilion, and V. Rizo, "Digital Filter for PCM Encoded Signals," US Patent 3,777,130, December 3, 1973.
12. A. Peled and B. Liu, "A New Hardware Realization of Digital Filters," IEEE Trans. Acoust., Speech, Sig. Proc., *ASSP-22* (December 1974), pp. 456–462.
13. R. E. Crochiere, S. A. Webber, and J. L. Flanagan, "Digital Coding of Speech in Subbands," B.S.T.J., *55,* No. 8 (October 1976), pp. 1069–1085.
14. J. Max, "Quantization for Minimum Distortion," Trans. IRE, *IT-6* (March 1960), pp. 7–12.
15. L. H. Rosenthal, L. R. Rabiner, R. W. Schafer, P. Cummiskey, and J. L. Flanagan, "A Multiline Computer Voice Response System Utilizing ADPCM Coded Speech," IEEE Trans. ASSP, *ASSP-22* (October 1974), pp. 339–352.

Contributors to This Issue

Anthony S. Acampora, B.S.E.E., 1968, M.S.E.E., 1970, Ph.D., 1973, Polytechnic Institute of Brooklyn; Bell Laboratories, 1968—. At Bell Laboratories, Mr. Acampora initially worked in the fields of high power microwave transmitters and radar system studies and signal processing. Since 1974, he has been studying high capacity digital satellite systems. His current research interests are modulation and coding theory, time diversion multiple access methods, and efficient frequency re-use techniques. Member, Eta Kappa Nu, Sigma Xi, and IEEE.

William C. Ahern, B.S.E.E., 1956, Newark College of Engineering; M.S.E.E., 1961, New York University; Bell Laboratories, 1959–1977. Mr. Ahern first worked on air defense and communication systems. In 1966 he became supervisor of a network planning group. From 1968 to 1971 he supervised a government network simulation and analysis group. From 1971 until his death in 1977, he supervised surveys of loop signal power and quality of network service.

Corrado Dragone, Laurea in E.E., 1961, Padua University (Italy); Libera Docenza, 1968, Ministero della Pubblica Istruzione (Italy); Bell Laboratories, 1961—. Mr. Dragone has been engaged in experimental and theoretical work on microwave antennas and solid-state power sources. He is currently concerned with problems involving electromagnetic wave propagation and microwave antennas.

Francis P. Duffy, B.A., 1965, King's College; M.S., 1968, Stevens Institute of Technology; Bell Laboratories, 1965—. Mr. Duffy has been involved in conducting statistical surveys to determine telephone network performance and customer behavior characteristics. Currently, he is involved in studying customer trouble reports.

David D. Falconer, B.A.Sc., 1962, University of Toronto; S.M., 1963, and Ph.D., 1967, Massachusetts Institute of Technology; post-doctoral research, Royal Institute of Technology, Stockholm, 1966–1967; Bell

Laboratories, 1967—. Mr. Falconer has worked on problems in communication theory, and high-speed data communication. During 1976–77 he was a visiting professor of electrical engineering at Linköping University, Linköping, Sweden. He presently supervises a group working on digital signal processing for speech bit rate reduction. Member, Tau Beta Pi, Sigma Xi, IEEE.

G. S. Fang, B.S.E.E., 1967, National Taiwan University; Ph.D. (E.E.), 1971, Princeton University; Computer Sciences Corporation, 1971–72; Bell Laboratories, 1972–1977. At Bell Laboratories, Mr. Fang worked on high-speed digital transmission, protection switching, and microprocessor applications. Since the fall of 1977, he has been with the Department of Electrical Engineering, National Taiwan University, Taipei, Republic of China.

Ben Gotz, B.E.E., 1966, The City College of New York; M.E.E., 1968, and Ph.D., 1971, New York University; Bell Laboratories, 1966–1969, 1971—. Mr. Gotz is engaged in studies related to speech coding for bit rate compression.

James A. Maher, B.S., 1962, Seton Hall University; M.S., 1969, Ph.D., 1973 (Applied and Mathematical Statistics), Rutgers University; instructor, Rutgers University, 1970–1973; Bell Laboratories, 1973—. Mr. Maher has been involved in planning statistical surveys designed to characterize network performance. He is presently involved in studies dealing with characterizing customer expectations and perceptions of network service. Member, ASA, ASEE, and Sigma Xi.

E. J. Messerli, B.A. Sc. (E.E.) 1965, University of British Columbia; M.S. (E.E.) 1966, Ph.D. (E.E. and C.S.) 1968, University of California, Berkeley; E.E. & C.S. faculty, Berkeley, 1968–69; Bell Laboratories, 1969—. Mr. Messerli has been primarily involved in systems analysis and network planning. His work includes studies on the demand assignment of capacity for a domestic satellite system, on the impact of faulty trunks on customers and the network, and on the worth of more accurate data for trunk provisioning. He is currently supervisor of a group concerned with planning for the integrated measurement of network performance; Member, IEEE, ORSA.

Debasis Mitra, B.Sc. (E.E.), 1964, and Ph.D. (E.E.), 1967, London University; Bell Laboratories, 1968—. Mr. Mitra has worked on stability analysis on nonlinear systems, semiconductor networks, computer memory management, analysis of queues in communication systems, growth models for new communication services, and adaptive systems. Most recently he has been involved in the analysis of speech coders and digital filters. Member, IEEE and SIAM.

V. Ramaswamy, B.Sc., 1957, Madras University, India; D.M.I.T., Madras Institute of Technology, Chromepet, Madras, India; M.S., 1962, and Ph.D., 1969, Northwestern University; Zenith Radio Corporation, 1962–65; Bell Laboratories, 1969—. Mr. Ramaswamy's previous work has included microwave components, diode parametric amplifiers and wave propagation in semiconductor plasmas. His present research interests are thin-film optical waveguides and devices and polarization effects in single mode optical fibers. Member, Sigma Xi, IEEE.

R. D. Standley, B.S., 1957, University of Illinois; M.S., 1960, Rutgers University; Ph.D., 1966, Illinois Institute of Technology; USASRDL, Ft. Monmouth, N.J., 1957–1960; IIT Research Institute, Chicago, 1960–1966; Bell Laboratories, 1966—. Mr. Standley has been engaged in research projects involving microwave, millimeter wave, and optical components. He is presently concerned with electron beam lithography as applied to fabrication of integrated optic devices. Member, IEEE, Sigma Tau, Sigma Xi.

THE BELL SYSTEM TECHNICAL JOURNAL is abstracted or indexed by *Abstract Journal in Earthquake Engineering, Applied Mechanics Review, Applied Science & Technology Index, Chemical Abstracts, Computer Abstracts, Computer & Control Abstracts, Current Contents/Engineering, Technology & Applied Sciences, Current Contents/Physical & Chemical Sciences, Current Index to Statistics, Current Papers in Electrical & Electronic Engineering, Current Papers on Computers & Control, Electrical & Electronic Abstracts, Electronics & Communications Abstracts Journal, The Engineering Index, International Aerospace Abstracts, Journal of Current Laser Abstracts, Language and Language Behavior Abstracts, Mathematical Reviews, Metals Abstracts, Science Abstracts, Science Citation Index,* and *Solid State Abstracts Journal.* Reproductions of the Journal by years are available in microform from University Microfilms, 300 N. Zeeb Road, Ann Arbor, Michigan 48106.

Bell System