THE BELL SYSTEM TECHNICAL JOURNAL

SEPTEMBER 1981
VOL. 60, NO. 7, PART 1

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, C. L. Brown, Chairman and Chief Executive Officer; W. M. Ellinghaus, President; V. A. Dwyer, Vice President and Treasurer; F. A. Hutson, Jr., Secretary. Editorial inquiries should be addressed to the Editor, The Bell System Technical Journal, Bell Laboratories, Room WB 1L-331, Crawfords Corner Road, Holmdel, N.J. 07733. 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.

ISSN0005-8580

Comments on the technical content of any article or brief are welcome. These and other editorial inquiries should be addressed to the Editor, The Bell System Technical Journal, Bell Laboratories, Room WB 1L-331, Crawfords Corner Road, Holmdel, N.J. 07733. Comments and inquiries, whether or not published, shall not be regarded as confidential or otherwise restricted in use and will become the property of the American Telephone and Telegraph Company. Comments selected for publication may be edited for brevity, subject to author approval.

THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 60 September 1981 Number 7, Part 1

Sampling From Structured Populations: Some Issues and Answers

By V. N. NAIR and T. E. DALENIUS*

(Manuscript received February 25, 1981)

This paper reviews some sampling issues that are common to many Bell System surveys. We discuss various aspects of two-stage sampling designs, and emphasize sampling from populations with multiple characteristics. The hierarchical structure of the population in many surveys makes the use of multistage sampling techniques attractive. In populations with multiple characteristics, often not every characteristic is common to every unit. We consider some special designs for sampling from such populations. Finally, we discuss some issues in network sampling. Two recent Bell System surveys are used to illustrate most of the ideas discussed. One of the surveys deals with the estimation of traffic characteristics for various classes of service, while the other one is a survey of baseband transmission impairments.

I. INTRODUCTION

Sample surveys have played an increasingly important role in the Bell System in recent years as a means of providing an objective basis for decision making. To an extent, this has been due to the growing awareness among users of the survey results that, in most surveys, sampling is not the only source of error and often not the primary source. Even if a presumably complete census were taken instead of a sample, serious errors might exist in the results arising from various causes such as measurement or response errors.

* Brown University.

The growth in numbers, in recent years, has also been accompanied by a widening of the range (both in type and complexity) of the surveys. For many of these surveys, a simple and readily available sampling design can easily be adapted to the needs of the prevailing situation. More often, however, the problem at hand is sufficiently complex and nonstandard so that various parts of existing sampling theory have to be modified and pieced together to arrive at a reasonable solution.

Nevertheless, some sampling issues are common to a number of Bell System surveys. Most of these surveys involve sampling from populations that are highly structured, and any cost-efficient sampling design must take this structure into account. In this paper, we review some sampling issues that arose in two surveys currently under implementation. Both surveys possess some common features as well as features unique to themselves. Since these features are common to a large number of other surveys, an exposition of both the theoretical and practical considerations involved may prove beneficial to other survey practitioners. Let us first consider the two examples.

Example 1. Cost of service traffic usage studies (COSTUS)

The various Bell operating telephone companies (OTCS) carry out these surveys periodically to obtain an objective basis for distributing the traffic-sensitive costs for a jurisdiction, typically a state within an OTC, among its various classes of telephone service. Measurements of three traffic characteristics (busy-hour CCS, busy-hour peg count and 14-day peg count) from the sampled telephone lines are used to calculate the relative magnitudes of the traffic characteristics for each class of service. [CCS is a traditional unit for measuring the usage of channels (it stands for hundred call seconds per hour). Peg count is the number of calls actually handled.] These values are then used as inputs to the "embedded direct costs" analysis, which allocates most traffic-sensitive investments and expenses among the various classes of service.

The elementary units in this study are telephone lines corresponding to the various classes of service. These units, however, are clustered into central offices. In fact, each central office has a number of clusters associated with it, one cluster for each class of service. A reasonably cost-efficient design should take this hierarchical clustering into account, since the major portion of the costs in observing a line arises from visiting the central office and setting up the measuring equipment. Thus, a two-stage sampling design with central offices serving as primary sampling units (PSUS) and telephone lines serving as secondary sampling units (SSUS) seems attractive. This is even more so since the central offices provide service in a number of classes of service so that from each sampled central office, we can further subsample telephone lines from all the available classes of service.

Hence, COSTUS are examples of the use of a two-stage sampling design for a population with multiple characteristics. The different characteristics here correspond to the different classes of service. The parameters of the sampling design in COSTUS are determined so that the busy-hour CCS parameter for each class is estimated with a prescribed accuracy. One additional complication in these studies is the fact that not all central offices provide service in every available class. In some jurisdictions, there are some classes of service (such as coin) that are provided in only a few offices. The sampling literature refers to this as the problem of "partial variate pattern" (PVP). The presence of PVP causes difficulties in selecting an appropriate sample of central offices for the estimation of the parameters of all the classes of service.

Example 2. Survey of baseband transmission impairments

The aim of this survey, currently under development at Bell Laboratories, is to measure baseband transmission impairments for various trunk facility types. From each sampled trunk, estimates of various impairment characteristics, such as signal to C-notched noise ratio (s/n) and second- and third-order harmonic distortion (R2 and R3) are to be obtained. Although the near (transmitting) and far (receiving) end-drop equipment, in addition to the carrier system, determines the trunk type, it is known from past experience that the contribution from the carrier system is the dominant factor. Thus, we do not consider the influence of the end-drop equipment in this study. Six different measurement characteristics are to be measured from each sampled trunk and the parameters of seven different trunk types are to be estimated.

The elementary unit in this survey is the trunk. While the trunks are again clustered into central offices, this clustering is not unique since one trunk is common to a pair (transmitting and receiving) of central offices. In fact, the structure of the population here resembles a graph (network) with the central offices as nodes and trunks as edges (arcs). This survey is an example of network (graph) sampling (see Ref. 1, for example). In this survey, if we sample a particular trunk, we have to visit the pair of end offices connected to the trunk to set up the measuring equipment. This implies that it is cheaper to sample additional trunks connected to those two end offices. Hence, taking the structure of the population into account results in considerable cost savings.

One possible approach to this problem is to use multistage sampling to select pairs of offices and trunks connected to those offices. Since we are interested in different trunk types, this study also involves multiple characteristics.

Both the above examples involve using multistage sampling to study populations with multiple characteristics. Multistage sampling is not an uncommon phenomenon in Bell System surveys where the natural administrative and geographic clustering of units makes it very cost efficient. In Sections II and III we review various issues that confront a survey statistician in developing a two-stage sampling design for studying multiple characteristics. Some of the issues discussed in Section II are also common to other sampling designs. Section III deals primarily with determining the parameters of the sample design. In Section IV, we consider some sampling designs for populations with PVP. Section V is a brief review of issues in network sampling. We conclude the paper with a summary in Section VI. Throughout the paper we try to balance theoretical considerations with practical guidelines gained from our own experience. One of the two examples is used, wherever possible, to illustrate the ideas discussed.

II. TWO-STAGE SAMPLING: SOME PRELIMINARIES

This section deals with some preliminary considerations in developing a two-stage sampling design. Some of the discussion deals with issues that are common to sample surveys in general. We begin with a discussion of the rationale for using two-stage or multistage sampling designs. After an introduction to some notation, we examine how prescribed accuracy requirements are implemented in a sample survey and discuss the use of prior information. Section 2.6 examines the use of varying probability sampling schemes. Section 2.7 discusses ratio estimators with specific emphasis on two-stage sampling situations.

2.1 Why two-stage sampling?

The individuals whose characteristics are to be measured in a study are called elementary units. Observational access to the elementary units, in many cases, is provided by multistage sampling. Let the elementary units be grouped into a number of suitable clusters. In two-stage sampling, the clusters are used as PSUs and a sample of PSUs is selected in the first stage. The PSUs selected are divided into a number of SSUs and a sample of SSUs is selected from each PSU selected in the first stage. (The elementary units themselves can serve as SSUs.) All elementary units in the selected SSUs are observed with respect to the variables of interest.

There are various reasons why multistage sampling is attractive. For instance, in many studies, a complete list ("frame") of elementary units is not available and it may be prohibitively expensive to create such a list. If it is relatively cheap to construct a list of clusters, the clusters can be used as PSUs in a two-stage sampling scheme. Then, only a list of the elementary units in the sampled clusters needs to be constructed. This results in considerable cost savings.

Often, the population of elementary units in a survey is dispersed over a large geographical area. If we have to visit each sampled unit to collect measurements, sampling from the list of elementary units can lead to high costs per elementary unit. A more cost-efficient scheme may be obtained by grouping the elementary units into geographically compact clusters and using multistage sampling with the clusters as PSUs.

Typically, the cost reduction in multistage sampling is accompanied by an increase in the variance of the estimate over the variance of an estimate from a simple random sampling (SRS) of the same number of elementary units. However, the "accuracy" per unit cost may be higher. If we have some control over the formation of the clusters, we can actually reduce the variance (relative to SRS) by grouping the units so that there is more variation within clusters than between clusters. In most Bell System surveys, however, the clusters are predetermined.

2.2 *Notation*

We use the following notation throughout the remainder of this paper:

M = number of PSUs in the universe,
m = number of PSUs sampled,
N_i = number of SSUs in PSU i, $i = 1, \cdots, M$,
n_i = number of SSUs selected from the ith sampled PSU, $i = 1, \cdots, m$,
Π_i = probability of selecting the ith PSU in a sample of size m, $\sum_{i=1}^{M} \Pi_i = m$,
Y_{ij} = characteristic to be measured, $j = 1, \cdots, N_i$, $i = 1, \cdots, M$,
y_{ij} = value corresponding to a sample unit, $j = 1, \cdots, n_i$, $i = 1, \cdots, m$,

$$Y_i = \sum_{j=1}^{N_i} Y_{ij} \qquad Y = \sum_{i=1}^{M} Y_i, \qquad \bar{\bar{Y}}_i = Y_i/N_i,$$

$$\bar{\bar{Y}} = Y/N, \qquad N = \sum_{i=1}^{M} N_i, \qquad S_i^2 = \frac{1}{N_i} \sum_{j=1}^{N_i} (Y_{ij} - \bar{\bar{Y}}_i)^2,$$

$$y_i = \sum_{j=1}^{n_i} y_{ij}, \qquad y = \sum_{i=1}^{m} y_i, \qquad \bar{\bar{y}}_i = y_i/n_i,$$

$$\bar{\bar{y}} = y/n, \qquad n = \sum_{i=1}^{m} n_i.$$

We consider only equal probability sampling schemes in stage two in this paper. The parameter of interest is the overall total $Y = \sum_{i=1}^{M} N_i \bar{\bar{Y}}_i$. $\hat{Y}$ denotes an arbitrary estimator of Y. The same considerations can be used for estimating the average $\bar{\bar{Y}}$ if we rewrite

$$\bar{\bar{Y}} = \sum_{i=1}^{M} W_i \bar{\bar{Y}}_i, \qquad W_i = N_i/N.$$

2.3 Accuracy requirements

The sampling design in a carefully planned survey is determined so that either (*i*) the total cost of the survey is minimized subject to a prescribed requirement on the accuracy of the estimators or (*ii*) the accuracy of the estimators is maximized subject to a constraint on the cost. Since both approaches involve essentially the same considerations (see Section III), let us consider in some detail just the problem of minimizing cost subject to accuracy requirements.

A sampling design, where the units are randomly selected according to given probabilities of selection, permits us to make quantitative statements about the error involved in the estimators. This in turn allows us to determine the sample sizes so that the prescribed accuracy requirements are met. These requirements are typically stated in terms of the error $e = \hat{Y} - Y$ or some function of the error, $f(e)$, such as relative error, and can be expressed as

$$\Pr\{|f(e)| \le \delta\} \ge 1 - \alpha \tag{1}$$

for some constants α and δ. In COSTUS, for instance, the sample sizes are determined so that the absolute values of the relative error is less than or equal to 0.1 with probability at least 0.9, i.e., $\alpha = \delta = 0.1$. To implement the accuracy condition (1), large-sample theory is usually used to claim that Y is approximately normally distributed. (It is beyond the scope of this paper to discuss the adequacy of this normal approximation. The interested reader is referred to Refs. 2 to 5.) Equation (1) is equivalent to an expression of an upper bound on the variance [or mean-square error (mse) if $\hat{Y}$ is biased] of $\hat{Y}$.

When estimating several parameters, as in populations with multiple characteristics, we may require that several accuracy criteria be satisfied simultaneously. By using normal approximations, we can state this problem, in general, as minimizing the total cost of the survey subject to a constraint on the variances (or mse's) of the form

$$A\mathbf{v} \le \gamma, \tag{2}$$

where $\mathbf{v} = (v_1, \cdots, v_p)^T$ is the vector of variances (mse's) of the p estimators, A is a $k \times p$ matrix that specifies the k specific linear combinations of the variances that have to meet the accuracy conditions, and $\gamma = (\gamma_1, \cdots, \gamma_k)^T$ represents the bounds on the accuracies.

For example, if $k = p$ and A is the identity matrix, then all the p parameters need to be estimated with prescribed accuracy. If $k = 1$, then only one particular linear combination of the variances is needed to satisfy an accuracy criterion.

2.4 Variance components

Since the accuracy specifications can be stated in terms of the variances of the individual estimators, we need to examine the components of the variance of the estimator in a two-stage sampling scheme. This will aid us later (Section III) in determining the relative contributions to the variance from stages one and two and the tradeoffs in increasing the sample size in stage one versus that in stage two. If we restrict our attention to linear estimators of the form $\hat{Y}_\alpha = \sum_i \alpha_i \bar{\bar{y}}_i$ for estimating Y, we see that α_i must equal N_i/Π_i for the estimator to be unbiased. With this choice of α, $\hat{Y}_\alpha$ is the well-known Horvitz–Thompson (H–T) estimator.[6] A discussion of some of the properties of this estimator can be found in Ref. 7. Let us restrict our attention to the H–T estimator and examine its variance.

If we select m PSUs with replacement (WR) in stage one with inclusion probabilities Π_i, we have a multinomial sample of size m with success probabilities $Z_i = \Pi_i/m$. If the second-stage units are chosen without replacement, the variance of

$$\hat{Y} = \frac{1}{m} \sum_{i=1}^{m} \frac{N_i}{Z_i} \bar{\bar{y}}_i$$

can be written as the sum of two components:[7–9]

(*i*) the within-PSU variation W is

$$W = \frac{1}{m} \sum_{i=1}^{M} \frac{N_i^2}{Z_i} \frac{S_i^2}{n_i} (1 - f_{2i}), \tag{3}$$

and

(*ii*) the between-PSU variation B is

$$B = \frac{1}{m} \sum_{i=1}^{M} Z_i (Y_i/Z_i - Y)^2.$$

Here,

$$S_i^2 = \frac{1}{N_i} \sum_{j=1}^{N_i} (Y_{ij} - \bar{\bar{Y}}_i)^2,$$

the within cluster variance and $1 - f_{2i} = (N_i - n_i)/N_i$, the finite population correction.

If the sampling is done without replacement (WOR) in stage one with varying selection probabilities, the within-PSU variation remains the same. The between-PSU variation, however, depends on second-order inclusion probabilities which are extremely hard to calculate.[7,10] Hartley and Rao provide some approximations.[10] One possible approximation is, of course, the use of eq. (3), valid for the WR scheme, in the WOR situation. If the sampling fraction m/M is large (say >0.25), this approximation may be unreasonable. When the sampling is done WOR with equal selection probabilities in stage one, i.e., SRSWOR, the B component is given by

$$B = \frac{M^2(1-f)}{m(M-1)} \sum_{i=1}^{M} (Y_i - \bar{Y})^2,$$

where $\bar{Y} = \frac{1}{M} \sum_{i=1}^{M} Y_i$ and $f = m/M$.

For a discussion of variance estimation in two-stage sampling, see Refs. 7, 8, or 9, for example. Some approximate but "quick and easy" methods of variance estimation are discussed in Refs. 11 and 12. If the variance estimator is intended only to provide a rough guide as to the accuracy of the estimator, an approximate, but quick and easy, method is adequate. If the accuracy of the estimator is of great importance and must be demonstrated through the variance estimator, we have to use a "good" variance estimator, such as one with small mse.

2.5 Prior information

We need prior information on the variance of the various estimators and on the sampling costs to determine the sample sizes in a survey. It is rare that we have very good prior information, particularly concerning the variance of the estimators. Preliminary estimates can be obtained from prior surveys or pilot studies. One practice commonly found in the Bell System is the use of data from the entire Bell System to develop preliminary estimates for specific jurisdictions.

To implement the accuracy conditions exactly in a two-stage sampling scheme, we need to know each one of the components of W and B in eq. (3) exactly. Since this is rather unlikely, we usually just use two numbers, one for W and one for B, instead of the individual values for each PSU. These numbers can be interpreted as either the average or the maximum over all PSUs.

When the quality of the prior information is poor (as a consequence of one or more of the above reasons), little can be gained in developing a complex design that may (or may not) be "optimum" for the problem at hand. A simpler design which is less sensitive to the preliminary estimates of the design parameters is more desirable. Also, when the preliminary variance estimators are unreliable, an estimate of the accuracy achieved should always be calculated after the fact from the sample to compare with the prescribed accuracy.

2.6 Varying probability sampling

The sample selection schemes in stages one and two can be based on equal or varying probability sampling techniques. For simplicity, we consider varying probability sampling only in stage one. The considerations here also carry over to other stages. Let us examine how the selection probabilities $\{\Pi_i\}$ should be determined so that the variance of the H–T estimator $\sum_{i=1}^{m} (N_i/\Pi_i)\bar{\bar{y}}_i$, for estimating $Y = \sum_{i=1}^{M} Y_i$, is minimized.

In the simpler situation of one-stage cluster sampling, i.e., $n_i = N_i$, if we take Π_i proportional to Y_i, the variance of the H–T estimator is zero.[7] Hence, if there exists an auxiliary variable X_i which is approximately proportional to Y_i, we can use this auxiliary information to select the Π_i's. In some two-stage sampling situations, we can use the measures of size of the PSU, $\{N_i\}$, to obtain "optimal" selection probabilities. To see this, note that the parameter Y can be written as $\sum_{i=1}^{M} N_i\bar{\bar{Y}}_i$, where $\bar{\bar{Y}}_i$ is the PSU mean, and often the $\bar{\bar{Y}}_i$'s are roughly of the same order of magnitude. In this case, the $Y = N_i\bar{\bar{Y}}_i$ will be roughly proportional to the N_i so that we can take the Π_i proportional to the N_i. This is known as probability proportional to size (PPS) sampling. (In COSTUS, for example, a priori, we expect the average busy-hour CCS per main station to be about the same across central offices.) When sampling from populations with multiple characteristics, there are multiple measures of size, one associated with each characteristic. The optimal selection probabilities are some function of these size measures, depending on the particular accuracy criteria of interest. In addition, there are also cases in which the exact size measures are unknown and we have to use estimated measures.

To develop a cost-efficient design, we need to minimize variance per unit cost rather than the actual variance. The optimal selection probabilities must therefore take the cost structure into account. In COSTUS, where the PSUs are central offices, the sampling costs depend on the type of switching equipment in the office. For example, it is considerably more expensive to visit and set up the measuring equipment in an electronic switching system (ESS) office than in a non-ESS office. If we use formal optimality calculations, we find that with other factors held constant, the optimal selection probability for each PSU is inversely proportional to the square root of the cost of sampling that PSU.[9]

One or more of the above considerations may indicate that even if the PSUs vary greatly in size, the optimal selection probabilities are not too unequal. In such a case, we may be better off using SRS, i.e., equal selection probabilities, since (*i*) the selection scheme is simpler and (*ii*) exact variance formulas are available if, in addition, we are sampling WOR. In some situations, we can actually calculate the gain from using varying selection probability schemes.[7] If the gain is not substantial in these situations, the use of SRS seems preferable.

Also, even if we use SRS when the PSUs vary greatly in size, we can use ratio estimators, which take into account this variation, to estimate the parameters. This is discussed in Section 2.7.

Finally, we briefly discuss a simple scheme for selecting PSUs with unequal probabilities. Many schemes for unequal probability selection exist,[3,7,10] and, in fact, several procedures may lead to the same inclusion probabilities $\{\Pi_i\}$. The scheme we consider here is for sampling WOR and is known as PPS systematic sampling. Let $\{T_i\}$ denote the cumulative totals of the desired selection probabilities $\{\Pi_i\}$,

$$\sum_{i=1}^{M} \Pi_i = m, \qquad T_i = \sum_{j=1}^{i} \Pi_j.$$

To select m PSUs, first select a random number $u \in [0, 1]$ and then select the m PSUs for which

$$T_{i-1} < u + j \leq T_i, \qquad j = 0, 1, \cdots, m - 1.$$

Hartley and Rao consider this procedure with a random arrangement of the PSUs and develop approximate variance expressions for the estimator.[10]

2.7 Use of ratio estimation

So far we have considered only unbiased estimators of the total Y. In some situations we can exploit information available for some auxiliary variable and use a biased estimator, such as the ratio estimator, which has smaller mse than the unbiased estimators. To see this, let $\{X_i\}$ be the known auxiliary variable and let X denote the total corresponding to this variable and $\hat{X}$ denote the estimator of X based on the sample. Since we know the error $\hat{X} - X$, we know how this sample performs in estimating X. Hence, if $\{X_i\}$ and $\{Y_i\}$ are highly correlated, it is intuitively clear that we can improve our original estimator $\hat{Y}$ by exploiting our knowledge of how well the sample estimates X.

The ratio estimator itself is a special case of the general difference estimator $\hat{Y}_a = \hat{Y} + a(\hat{X} - X)$ and is obtained by taking $a = -\hat{Y}/\hat{X}$. This results in the estimator $\tilde{Y} = (\hat{Y}/\hat{X})X$. There are other ways of exploiting the information about $\hat{X} - X$. For instance, a can be a prespecified constant. (If $a = 0$, we get the original estimator $\hat{Y}$ based on the Y measurements alone.) We can also take a to be the regression coefficient $\hat{\beta}$ obtained by regressing the Y_i's on the X_i's.

For the ratio estimator $\tilde{Y}$, the mse of Y can be approximated up to a first-order term by

$$V(\hat{Y}) - 2R\,\mathrm{Cov}(\hat{Y}, \hat{X}) + R^2V(\hat{X}),$$

where $R = Y/X$ and $\hat{Y}$ and $\hat{X}$ are unbiased estimators of Y and X.[9] Thus, $\tilde{Y}$ will be more efficient than the unbiased estimator $\hat{Y}$ if $2R\,\mathrm{Cov}(\hat{Y}, \hat{X}) > R^2V(\hat{X})$. This is likely to be true in practice if the X_i's are appropriately chosen.

In Section 2.6, we saw that in some two-stage sampling situations, the Y_i's are likely to be correlated with the size measures $\{N_i\}$. If PPS sampling is not used (for one or more of the reasons we considered earlier), we can take the N_i's to be the auxiliary variables and use the resulting ratio estimator $\tilde{Y} = \hat{R}N$, where

$$\hat{R} = \sum_{i=1}^{m} \frac{N_i}{\Pi_i}\bar{\bar{y}}_i \Big/ \sum_{i=1}^{m} \frac{N_i}{\Pi_i}.$$

(If we use PPS sampling, the ratio estimator with the size measure as the auxiliary variable is the same as the unbiased estimator.) Our experience with data from several jurisdictions for COSTUS showed a considerable gain from the use of this ratio estimator.

III. DETERMINING THE DESIGN PARAMETERS

3.1 Cost considerations

The ultimate objective in designing an efficient survey design is the maximization of accuracy per unit cost. To accomplish this, we need to know the cost structure of the survey. We can identify three types of costs in two-stage sampling: (*i*) overhead costs; (*ii*) costs that depend primarily on the number of PSUs in the sample; and (*iii*) costs that depend on the number of SSUs in the sample. Since the overhead costs are fixed, they can be ignored in determining the sample sizes. The costs of sampling PSUs may consist of the costs of selecting, traveling to, locating each sampled PSU, and setting up the measuring equipment. A simple cost function may be of the form

$$mC_1 + m\bar{n}C_2, \tag{4}$$

where C_1 and C_2 are the costs of sampling a PSU and SSU, respectively, and $m\bar{n}$ is the total number of SSUs sampled. Typically, however, the cost functions are more complex. In COSTUS, as we mentioned earlier, the cost of sampling a PSU varies from one PSU to another and depends primarily on the switching equipment in the central office. Further, the cost of sampling a telephone line (SSU) also depends on the switching equipment and so varies from one office to another. There is also a special cost structure in the transmission impairments survey in Example 2. Here, if trunks (edges) are selected by using two-stage sampling to determine the pair of end offices connected to the trunk, it is cost-efficient to select offices with many trunks rather than those with fewer trunks.

3.2 Determining the parameters in a simple situation

Let us consider a simple situation to illustrate the concepts involved in determining the parameters of the optimum design. Suppose the number of SSUs in each PSU is the same and equals $\bar{N}$, the cost function is given by eq. (4) and we use SRSWOR to select the units in both stages. We need to determine only m, the number of PSUs to be sampled, and $\bar{n}$, the number of SSUs to be sampled from each selected PSU. The variance of the H–T estimator can now be written (see Section 2.4) as

$$V(\hat{Y}) = (1 - f_1)\frac{V_1^2}{m} + (1 - f_2)\frac{V_2^2}{m\bar{n}} \tag{5}$$

for some V_1 and V_2. Here $f_1 = m/M$ and $f_2 = \bar{n}/\bar{N}$. A comparison of eq. (5) with the cost function $C = mC_1 + m\bar{n}C_2$ reveals that increases in m and $\bar{n}$ have opposite effects on the variance and costs. Also, it is clear that an increase in m results in greater reduction in the variance than a corresponding increase in $\bar{n}$. Since C_1 is typically much larger than C_2, it is more costly to increase the size of the first stage sample than the size of the second stage sample. All of these factors must be taken into consideration in determining the optimum combination of m and $\bar{n}$.

As mentioned earlier, optimum levels of m and $\bar{n}$ can be determined by minimizing either (*i*) the variance subject to a cost constraint or (*ii*) the cost subject to some accuracy requirements. Both approaches yield essentially the same results. The problem can be formulated mathematically as minimizing a given function subject to a constraint. Standard numerical or analytical techniques (LaGrangian multipliers, Cauchy's inequality) can be used to determine the optimum values of m and $\bar{n}$. In this particular simple situation, explicit expressions for m and $\bar{n}$ can be easily obtained. Suppose we want to minimize the cost subject to the condition that the variance eq. (5) does not exceed some value b. If we can ignore the finite population corrections in eq. (5), the optimum values of $\bar{n}$ and m can be obtained as

$$\bar{n}_{\text{opt}} = \frac{V_2/V_1}{\sqrt{C_2/C_1}}$$

and

$$m_{\text{opt}} = \frac{V_1/\sqrt{C_1}}{A},$$

where

$$A = b/(C_1/V_1^2 + C_2/V_2^2)^{1/2}.$$

The total cost of the survey with these values of m and n is given by

$$C_{\text{opt}} = b/A^2.$$

In practice, one should not be satisfied with just determining the optimum values of m and $\bar{n}$ without examining the behavior of the variance and cost functions near the optimum. Since preliminary estimates of costs and variances may only be approximate, the behavior of these functions in a neighborhood around the optimum should be examined. Relatively flat variance and cost functions near the optimum value indicate robustness against possible moderate errors in the input parameters.

3.3 More general situations and the COSTUS example

In most surveys, the situation is more complex than the one we have just discussed. For example, the PSUs will not necessarily be the same size and the cost function may be more complicated. Even in a general situation, the problem can be formulated in such a way that we can determine, either analytically or numerically, the optimum values of: m, the number of PSUs to be selected; $\{\Pi_i\}$, the inclusion probabilities; and $\{n_i\}$, the number of SSUs to be sampled from each selected PSU. Some of these results for some special cost functions can be found in the literature.[7-9]

We want to emphasize here the importance of simplifying the problem, whenever possible, by using reasonable approximations. In a complex situation where there are too many design parameters to be determined, it is difficult to appreciate the impact of unreliable input values. Reducing the number of parameters through the use of some practical guidelines usually provides us with a better understanding of the problem. We illustrate some of these ideas through the COSTUS example.

The PSUs in COSTUS are central offices and, as mentioned earlier, the cost of sampling the office and telephone lines (SSUs) in the office depends on the type of switching equipment in the office. Since each office provides service in several classes, we have to sample lines from all the available classes in the selected offices. However, not every office provides service in every available class. Since we want to study the parameters of all the classes, we take the first-stage costs of sampling an office with service in only one class to be twice that of an office with service in two classes. Hence, the total costs of the survey can be written as

$$TC = \sum_{i=1}^{m} \left\{ \tilde{D}_{1i} + \sum_{C=1}^{S} D_{2i} n_{Ci} \right\},$$

where $\tilde{D}_{1i} = D_{1i}/\Sigma_i$, D_{1i} = the costs of sampling the ith office, Σ_i = number of classes in the ith office, D_{2i} = costs of sampling a line from the ith office, and n_{Ci} = number of lines to be selected from the ith office for class C (equals zero if office i does not have service in class C). Since the total cost of this survey is a random quantity, we minimize the expected total cost

$$m\left[\sum_{i=1}^{M} Z_i\left(\tilde{D}_{1i} + D_{2i}\sum_{C=1}^{S} n_{Ci}\right)\right], \tag{6}$$

where $mZ_i = \Pi_i$, the inclusion probabilities.

We need to minimize eq. (6) subject to some accuracy constraints. In this study, the quantity to be estimated is the mean load (in CCS) during the busy hour, $\bar{\bar{Y}}_C$. We require a relative error no larger than 0.1 with probability 0.90 for each of the S classes, $C = 1, \cdots, S$. From Section 2.3, we note that this accuracy can be stated in terms of an upper bound on the mse of the estimator. We use the following approximate expression for the relative mse (rmse) to determine the design parameters:

$$\mathrm{rmse}(\hat{\bar{\bar{Y}}}_C) = \bar{\bar{Y}}_C^{-2}\left[\sum_{i=1}^{M}\frac{W_{Ci}^2}{mZ_i}\left(\frac{S_{Ci}^2}{n_{Ci}} + (\bar{\bar{Y}}_{Ci} - \bar{\bar{Y}}_C)^2\right)\right]. \tag{7}$$

The notation here is the same as in Section 2.2. The additional subscript indicates the class of service. This expression (which in fact equals the relative variance of the unbiased estimator) is the zeroth-order term in the Taylor series expansion for the rmse of the ratio estimator. By not taking into account the higher-order terms which include the correlation between the numerator and denominator of the ratio estimator, this expression, in general, overestimates the variability. However, it is simpler to use and the overestimation may be desirable in view of the unrealiability in preliminary estimates.

Before determining the design parameters, we make two additional simplifications: (i) replace $S_{Ci}^2/\bar{\bar{Y}}_C^2$ in the first component of eq. (7) by V_{C2}, a quantity that does not depend on the office i; and (ii) replace $(\bar{\bar{Y}}_{Ci} - \bar{\bar{Y}}_C)^2/\bar{\bar{Y}}_C^2$ in the second component by V_{C1}, also a quantity independent of the office. This is reasonable since a priori we do not expect much variation between these values and, in any event, we do not know each one of the individual values. (See also the discussion on prior information in Section 2.5.)

Thus, we want to determine the design parameters which minimize eq. (6) subject to

$$\sum_{i=1}^{M}\frac{W_{Ci}^2}{mZ_i}\left(V_{C1} + \frac{V_{C2}}{n_{Ci}}\right) \le b_C$$

for some b_C, $C = 1, \cdots, S$. Instead of determining m, $\{Z_i\}$ and $\{n_i\}$ from the optimality calculations, we only determine m and $\bar{n}_C$, the average number of SSUs to be sampled from a selected PSU for each class of service. Once m and $\bar{n}_C$ are determined, we can allocate $m\bar{n}_C$, the total number of lines for class C, to each sampled office inversely in proportion to $(D_{2i})^{1/2}$. We also select $\{Z_i\}$ in advance by taking them proportional to

$$\{N_{\cdot i}/\mathrm{D}_{1i}{}^{1/2}\},$$

where

$$N_{\cdot i} = \textstyle\sum_{C=1}^{S} N_{Ci}.$$

Once we substitute these values for $\{n_{Ci}\}$ and $\{Z_i\}$ in the variance and cost functions, it is a relatively easy problem to find the values of m and $\bar{n}_C$ that minimize the total expected cost subject to the accuracy constraints. Since there are only $S + 1$ design parameters involved, it is also easy to examine the behavior of the cost and variance functions near the optimum and investigate the sensitivity to errors in input values.

When COSTUS was implemented in a few jurisdictions, we also examined the advantage gained by using unequal probability selection schemes. Since we were using the conservative WR variance formulas for the unequal probability selection scheme, we found that the loss in "efficiency" from using SRSWOR of offices (with exact variance formulas) was not substantial. This also simplified the computations considerably.

IV. SAMPLING DESIGNS FOR POPULATIONS WITH PARTIAL VARIATE PATTERNS

4.1 The problem of partial variate pattern (PVP)

A multivariate population (for example, one with multiple characteristics) is said to exhibit a PVP if not all the variates can be observed from every unit in the population. In COSTUS, as we noted, not all the central offices provide service in every available class. In the survey of baseband transmission impairments in Example 2, not all carrier systems appear between each pair of central offices. It is easy to visualize many other studies, both within and outside the Bell System, where the populations exhibit PVP. The problem of PVP can be serious if there is great variation in the size of the universe corresponding to each variate. The usual sampling designs may not provide reasonable assurance that we can select a sample that will allow us to estimate the parameters corresponding to each variate with prescribed accuracy.

Let us consider some schemes for sampling in the presence of PVP (also see Ref. 13). Since the problem of PVP is present in one stage of the selection process only, we restrict our attention to sample selection in the first stage. Thus, suppose there are M units in the population, of which M_C units have characteristic C, $C = 1, \cdots, S$. Let the sample size, determined by accuracy requirements, for characteristic C be m_C. These sample sizes of course also depend on the particular sampling scheme used.

4.2 Some sampling designs

4.2.1 Modified simple multivariate sampling

Let $m = \max_C m_C$ and suppose we select a sample of $m < M$ units, possibly using different selection probabilities for different units. This is the simple multivariate sampling scheme, intended for populations with no PVP. If $\tilde{m}_C$ denotes the number of sampled units with characteristic C, $\tilde{m}_C$ may be much smaller than m_C and in some cases may even be zero. We can modify this scheme in a number of ways. Instead of selecting $m = \max_C m_C$ units, we can select m^* units, according to selection probabilities $\{\Pi_i\}$, where m^* is determined so that the expected number of units in the sample is at least m_C, $C = 1, \cdots, S$. This can be achieved by taking $m^* = \max_C m_C/p_C$, where p_C is the total of the probabilities $Z_i = \Pi_i/m$ for units with characteristic C. This can be justified if we view the selection of a unit with chracteristic C approximately as a binomial experiment with probability of success p_C. This formulation can alternatively be used to determine m^* such that, say 90 percent of the time, $\tilde{m}_C \geq m_C$, $C = 1, \cdots, S$.

4.2.2 Combined multivariate sampling

Here, we consider S universes, each universe corresponding to the units with characteristic C, $C = 1, \cdots, S$. We select an independent sample of size m_C from each one of the S universes. We then observe every available characteristic from the units selected in all of the S samples. The total number of units selected in these S samples can vary between $\max_C m_C$ and $\sum_{C=1}^{S} m_C$. The main disadvantage of this scheme is that this number may be too large. However, we can exercise some control over this number. One possibility is to give higher selection probabilities to units with more characteristics than those with fewer characteristics (see Section 3.3). Alternatively, instead of selecting m_C units from the universe corresponding to characteristic C, we can select a smaller number, m_C^*, of units. This is because we expect to select some units, in addition to these m_C^* units, with characteristic C from the remaining $S - 1$ samples. So, the number m_C^* can be determined such that either on the average or with prescribed probability, the total number of units with characteristic C exceeds m_C, $C = 1, \cdots, S$. The binomial approximations discussed earlier can be used to determine the $\{m_C^*\}$.

4.2.3 Stratified sampling

We can also try to deal with PVP by stratifying the units so that, within each stratum, the units are internally homogeneous in some sense in terms of the PVP. We consider two stratification techniques here.

In the first scheme, called variate stratification, the strata are determined in terms of the variates (characteristics). Suppose the variates are ordered so that the number of units with variate one is smallest, the number with variate two is next smallest, etc. Then, stratum one consists of all the units with variate one, stratum two consists of all units with variate two and not in stratum one, etc. If we now allocate the total sample size among the strata, we can estimate the parameters corresponding to all the variates, especially the "small" ones. However, this scheme is not foolproof in the sense that it is possible to construct examples where the selected sample does not contain any units with one of the variates.

The second method, pattern stratification, is based on the variate pattern. Here, units with identical variate pattern, i.e., having the same set of characteristics, are grouped into a stratum. Unlike the variate stratification scheme, we can guarantee the required sample size for each variate in this scheme. However, this scheme suffers from the serious drawback that the total sample size may be too large, since the number of different strata (which is smaller than the sample size) can be as large as $\min(M, 2^S - 1)$.

In both these schemes, standard nonlinear programming techniques can be used to determine the sample size for each stratum to minimize cost subject to the variance constraints.

4.2.4 Other methods

It is possible to use sequential sampling schemes to ensure that we select a sample with a given number of units for each characteristic.[13] However, it is extremely difficult to determine analytically the selection probabilities for most of these schemes. One simple sequential method that can be implemented is a two-stage simple multivariate sampling scheme in which a simple multivariate sample is supplemented by a second-stage sample from the remaining units. Although the variance calculations become more involved, they are still tractable.

It also is plausible that ideas from the controlled selection methodology can be applied to the selection of samples from populations with PVP.[14,15,16] However, it is not clear how to characterize explicitly the set of all feasible samples here. Variance calculations also remain a difficult problem with controlled selection.

4.3 The design used in COSTUS

The sampling design used in COSTUS for handling PVP will be described here. As the problem of PVP exists only in the first stage, we consider the selection of units in stage one only.

While examining data from several jurisdictions for the different PVPS, we found that, in most cases, a class of service can be classified as either small or large in terms of the proportion of offices with service in that class. There were very few jurisdictions with medium-sized classes of service.

Since the main concern in the presence of PVP is the ability to estimate parameters corresponding to the small classes of service, we decided to group all offices with services in these classes in stratum 1. A combined multivariate sampling scheme, which guarantees the required sample size from each class, is used to select offices from this stratum. Since the total number of offices sampled under this scheme may be large, we restrict the size of this stratum to be no larger than 25 percent of the universe.

We can use a simple multivariate sampling scheme to select a sample from the remaining offices. However, we first identify those classes with service in less than 50 percent of the remaining offices. The offices with service in these classes (and not in stratum 1) are grouped into stratum 2. The remaining offices are grouped into stratum 3. Simple multivariate sampling schemes are then used to select units in strata 2 and 3. By doing this, we have reasonable assurance that the sample sizes for the classes that characterize stratum 2 are not too small compared to the required sizes.

Hence, we see that the sampling design for COSTUS is in fact a three-stage sampling design. In the first stage, the offices are grouped into three strata. Different sampling schemes are used in the different strata to select offices in the second stage. From each office selected in the second stage, telephone lines corresponding to each available class are selected in the third stage.

The design we have used here for handling PVP incorporates specific features of some of the schemes discussed in Section 4.2. The stratification is based on considerations ,similar to those in the variate stratification scheme. It is, however, adaptive in the sense that it depends on the variate pattern in each universe. In our applications, we found that in many jurisdictions stratum 2 was empty and in some situations, where the problem of PVP is not serious, stratum 1 was empty.

We arrived at the final design used in COSTUS by examining data from various jurisdictions for the different types of PVP to expect. This design, while not foolproof, provides a reasonable, practical solution to the problem at hand.

V. SAMPLING FROM NETWORKS

In most surveys, we can treat the population under study merely as a collection of elementary units with no importance attributed to the interrelationships that exist among the units. In some situations, however, these relationships cannot be ignored and the selection of the sample is necessarily affected by the network of relationships that exist in the population. In this section, we briefly review some aspects of network sampling and discuss the sampling design used in Example 2.

5.1 Networks

There are a wide range of surveys in the Bell System that deal with sampling from a network. Besides communication networks, network sampling also occurs in studies of other types of traffic flow and transportation facilities. A contact network or sociogram may represent the interrelationships among a group of individuals, households, customers, etc. Other examples include similarity or dissimilarity structures in cluster analysis and multidimensional scaling, where we want to compare a set of objects and group them into classes of similar objects.

A network can be described in abstract terms with the aid of graph theory. An undirected graph (network) consists of a nonempty set V of elements called vertices (nodes) and a set of E of elements called edges. Each edge e of E is associated with a pair of vertices (i, j). The edges may have several attributes associated with them. A network can also be represented by a matrix with the columns and rows representing the vertices. A one in the (i, j)th cell of the matrix indicates that the vertices i and j are connected. In the survey of baseband transmission impairments discussed in Example 2, the vertices are central offices and the edges are trunks. In this case, there are many trunks and also different types of trunks between a pair of central offices. Several attributes, corresponding to the impairment characteristics, are associated with each trunk.

5.2 Some sampling schemes

The manner in which we have observational access to the elementary units is the key to developing a reasonable sampling design. If we have a "frame" of all the edges in the graph from which we can select a sample of units, the problem is essentially one in traditional sampling theory. If no such frame is available and the structure of the relationship between the nodes must be discovered and explored during the course of data collection, the sampling design problem is quite different. Even in cases in which a complete listing of the edges is available, as in Example 2, cost considerations may dictate that a sample of edges be selected by first sampling the nodes. Also, unlike traditional sampling where information about a unit can be obtained only by sampling and observing it, information about the relationship between several nodes may be obtained at any one of the nodes in network sampling.

The field of sampling from networks has been considered by only a few authors so far.[1,17-22] Most of the attention has been focused on surveys for which the structure of interrelationships is unknown and must be discovered. The references above deal mainly with estimating parameters that measure various aspects of these relationships.

Goodman proposed the "snowball" sampling scheme for selecting edges (or pairs of connected nodes).[20] In this procedure, the survey proceeds from an initial sample of nodes by obtaining information about other nodes to which they are connected. The next step is to add to the sample some or all of these connected nodes, obtaining data from them as well as information about still other nodes to which they are connected. In an s-stage k-name snowball sample, this process is repeated for s stages and at each stage, k other nodes connected to a node already in the sample are selected. Goodman studies this scheme in detail under the assumption that the initial sample is selected through binomial sampling.[20] He also considers the case in which the k nodes are selected randomly at each stage. See also Ref. 1.

To consider two other methods of network sampling, let us view the network as a matrix with the vertices corresponding to the columns and rows and the elements of the matrix corresponding to the edges. If we select a sample of nodes (rows/columns of the matrix), we can base our inference entirely on the sampled subnetwork that corresponds to the sampled rows and columns. This procedure (called subnetwork sampling) of selecting one or even several subsystems out of a number of subsystems is equivalent to traditional one-stage cluster sampling. It leaves open all questions about interrelationships between one cluster and another. In the partial network sampling scheme, we select a sample of nodes from the node set, and observe all the edges connected to one or more of the nodes in the sample. Estimation of the network characteristics using these two schemes is discussed in Refs. 1 and 18.

5.3 Survey of baseband transmission impairments

In this survey, there are a number of trunks of various types with each trunk associated with a pair of end offices. If we select a particular pair of end offices, it then becomes cheaper to select additional trunks from those trunks that terminate in either one of the two offices. This special cost structure implies that we need to select trunks (edges) by appropriately selecting offices (nodes) to which they are connected.

A multistage sampling scheme is used in this survey. A sample of primary offices, using probabilities proportional to some measure of size (the number of trunks), is selected in the first stage. A number of secondary offices are selected, again using probabilities proportional to some measure of size, from the set of offices connected to each of the primary offices. From every pair of end offices thus sampled, a number of trunks corresponding to each trunk type are selected using simple random sampling. The parameters of the sampling design (m, the number of primary offices, $\{m_i\}$, the number of secondary offices and $\{n_{ij}\}$, the number of trunks of a particular type) can all be determined so that the total survey cost is minimized subject to some accuracy criterion.

The two-stage sampling scheme used here to select the pair of end offices can also be viewed as a two-stage snowball sampling scheme. It is of course possible to use a k-stage snowball sample to select the offices. Optimality considerations relating to the number of stages and the sample size in a snowball sample have yet to be resolved.

VI. SUMMARY

We have reviewed various aspects of sampling from structured populations in this paper. The issues that have been selected for discussion, two-stage sampling from populations with multiple characteristics and sampling designs for populations with PVP and network sampling, are common to many Bell System surveys. Thus, we hope that an exposition of some of the theoretical and practical considerations involved in dealing with these situations will serve other survey practitioners. Throughout the paper we have tried to balance theoretical considerations with practical guidelines gained from our own experience. Two recent Bell System surveys are used to illustrate the ideas disscused.

VII. ACKNOWLEDGMENTS

We are grateful to R. E. Hausman and C. H. Morton for their kind permission to use the baseband transmission impairments survey as an example. We have also benefited from the earlier experiences of M. T. Chao and F. W. Nolte in the Cost of Service Traffic Usage Studies.

REFERENCES

1. O. Frank, "Survey Sampling in Graphs," J. Statist. Planning Inference, *1* (1977), pp. 235–64.
2. J. Hajék, "Asymptotic Theory of Rejective Sampling with Varying Probabilities from a Finite Population," Ann. Math. Statist., *35* (1964), pp. 1491–523.
3. J. N. K. Rao, "Sampling Designs Involving Unequal Probabilities of Selection and Robust Estimation of a Finite Population Total," in *Contributions to Survey Sampling and Applied Statistics*, edited by H. A. David, New York: Academic, 1978.

4. B. Rosén, "Asymptotic Theory for Successive Sampling with Varying Probabilities Without Replacement, I and II," Ann. Math. Statist., *43* (1972), pp. 373-97, 748-76.
5. B. Rosén, "Asymptotic Theory for Des Raj's Estimator, I and II," Scand. J. Statist., *1* (174), pp. 71-83, 135-44.
6. D. G . Horvitz and D. J. Thompson, "A Generalization of Sampling Without Replacement from a Finite Universe," J. Am. Statist. Assoc., *47* (1952), pp. 663-85.
7. M. N. Murthy, *Sampling Theory and Methods*, Calcutta: Statistical Publ. Soc., 1977.
8. W. G. Cochran, *Sampling Techniques*, New York: Wiley, 1977, 3rd ed.
9. M. H. Hansen, W. N. Hurwitz, and W. G. Madow, *Sample Survey Methods and Theory*, Vols. I and II, New York: Wiley, 1953.
10. H. O. Hartley and J. N. K. Rao, "Sampling with Unequal Probabilities Without Replacement," Ann. Math. Statist., *33* (1962), pp. 350-74.
11. D. R. Brillinger, "Approximate Estimation of the Standard Errors of Complex Statistics Based on Sample Surveys," New Zealand Statist., *II*, No. 2 (1976), pp. 35-41.
12. B. V. Shah, "Variance Estimates for Complex Statistics from Multistage Sample Surveys," in *Survey Sampling and Measurement*, edited by N. K. Namboodiri, New York: Academic, 1978.
13. T. E. Dalenius and O. Frank, "Sampling Populations with Partial Variate Patterns," Scand. J. Statist., *1* (1974), pp. 19-27.
14. R. S. Cochran, "Sampling in Two or More Dimensions," in *Contributions to Survey Sampling and Applied Statistics*, edited by H. A. David, New York: Academic, 1978.
15. L. A. Goodman and L. Kish, "Controlled Selection—A Technique in Probability Sampling," J. Am. Statist. Assoc., *45* (1950), pp. 330-72.
16. R. J. Jessen, "Probability Sampling with Marginal Constraints," J. Am. Statist. Assoc., *65* (1970), 776-96.
17. A. R. Bloemena, *Sampling from a Graph*, Amsterdam: Mathematisch Centrum, 1964.
18. O. Frank, *Statistical Inference in Graphs,* Stockholm: Försvarets Forskningsanstalt, 1971.
19. O. Frank, "A Note on Bernoulli Sampling in Graphs and Horvitz-Thompson Estimation," Scand. J. Statist., *4* (1977), pp. 178-80.
20. L. A. Goodman, "Snowball Sampling," Ann. Math. Statist., *32* (1961), pp. 148-70.
21. M. Granovetter, "Network Sampling: Some First Steps," Am. J. Sociol., *81* (1976), pp. 1287-302.
22. F. F. Stephan, "Three Extensions of Sample Survey Technique: Hybrid, Nexus, and Graduated Sampling," in *New Developments in Survey Sampling*, edited by N. L. Johnson and H. Smith, New York: Wiley, 1969.

Copyright © 1981 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 60, No. 7, September 1981
Printed in U.S.A.

A First-Come-First-Serve Bus-Allocation Scheme Using Ticket Assignments

By D. K. SHARMA and S. R. AHUJA

(Manuscript received November 13, 1980)

This paper describes a new scheme for allocating a data bus on a first-come-first-serve (FCFS) basis. When the devices connected to the bus request to become the bus-master, they are assigned distinct "ticket numbers" in the order in which the requests are generated, at which time they go into a wait state. When the bus is released by a device holding the ticket number n, it is then allocated to the device holding the ticket number n + 1. We discuss the conditions under which the scheme is a close approximation to the ideal FCFS scheme and evaluate its performance using simulation results. We also present two alternative hardware implementations of this scheme—one centralized and the other distributed. Because of its simple hardware implementation, the scheme is attractive for applications where a bus is shared, in an unbiased fashion, among a large number of devices.

I. INTRODUCTION

In computer systems, situations frequently arise where a resource is shared among several devices, but it can be used by only one device at a time. Scheduling such a resource to enforce mutual exclusion over its use is necessary if devices request the resource while it is being used or if the requests arrive simultaneously. A frequently encountered resource of this type is the data bus, which provides a communication path among the various devices connected to it. At any given time, there can be several devices receiving (or reading) information from the bus, but there can be only one device that has the privilege of transmitting information on it. Such a device is called the bus-master, and mutual exclusion among the devices wishing to become the bus-master is enforced by bus arbitration schemes.

Devices requesting the bus while it is busy are made to wait until it becomes available again. As soon as that happens, one among the waiting devices is allowed to become the bus-master. In most bus arbitration schemes, this choice is made without regard to the order in which the requests originally arrived; for example, daisy-chaining, device polling, and parallel priority resolution schemes.

Some of the commonly used bus arbitration schemes have been reviewed by Chen and Thurber et al.[1,2] Among them, polling and daisy-chaining are most commonly used. Polling is suitable only for slow devices, because the waiting times from bus request to bus grant are quite long, as the devices can access the bus only during preassigned time intervals. Daisy-chaining is extensively used in several minicomputers, such as the PDP-11s made by Digital Equipment Corporation (DEC).[3] Arbitration delay in this scheme may be quite long, since it is proportional to the number of devices connected to the bus. Furthermore, by virtue of their location on the bus, the devices are assigned fixed priorities that are used for contention resolution.

For faster bus arbitration, the recent computers designed by DEC and Honeywell, Inc. use distributed schemes.[4,5] These schemes, and those described in Refs. 6, 7, and 8, use the same algorithm with different implementations. They are fast, modular, and flexible, but they, too, allocate fixed priorities to the devices connected to the bus.

The major drawback of allocating fixed priorities to the devices is that the low priorities may have to wait indefinitely before being granted access to the bus if a few high-priority devices decide to use the bus frequently. They are effectivelly "locked out" from service. See Ref. 9 for a simulation-based quantitative analysis of these and other bus arbitration schemes.

In this paper, we present a first-come-first-serve (FCFS) scheme that allocates the bus in an order that is a close approximation to that in which the devices request the bus. This scheme does not have the above-mentioned drawback of locking out a few devices from service, and it provides an equal grade of service to all devices. We first describe the scheme and then discuss two alternative hardware implementations—one centralized and the other distributed. Before describing the scheme, we briefly discuss the advantages of following the FCFS allocation policy.

Consider a data bus that is shared among several devices, and assume that (*i*) the devices request the bus with the same statistics, and (*ii*) the bus is allocated for a fixed quantum of time for each request. The bus arbitration scheme should then have the following two properties:

(*i*) It will minimize the idle time on the bus, so that the bus throughput is maximized. This is done by arbitrating for the next bus-master concurrently with the bus usage and by reducing the arbitration time, if it happens to be longer than the time quantum for which the bus is allocated.

(*ii*) It will have the least disparity of service across requests and also across devices. The disparity of service across requests is represented by s, the standard deviation of the waiting times taken over all bus requests, and the disparity of service across devices is represented by S, the standard deviation of the average waiting times experienced by the individual devices. It is a simple matter to show that if an arbitration scheme does not prefer a device over any other, the average waiting times experienced by individual devices are all equal. Such schemes are called unbiased schemes, and S for them is zero. The ideal FCFS scheme is one such scheme. In the Appendix we show that the ideal FCFS scheme also attains the minimum value of s. Thus, under the assumptions stated above, the ideal FCFS scheme is a desirable scheme to be emulated in real systems.

II. DESCRIPTION OF THE SCHEME

Let there be N devices connected to the bus. In order to ensure that the devices gain bus control one at a time and in the order in which they requested it, we propose a scheme that is very similar in essence to that used in many supermarkets. As customers walk in, they pull out a numbered ticket from a machine that dispenses sequentially numbered tickets. When the server becomes free, he or she waits on the customer with the ticket one number higher than that of the last customer served, thus, providing equitable service to all customers.

Our scheme is based upon two essential pieces of information: "next number to be served" (NNS) and "next number available" (NNA). This information can be maintained in a centralized or distributed fashion, as we discuss in Section III. In addition, each device has a register, called the ticket register, to store the ticket number assigned to it when it requests bus mastership. How these ticket numbers are assigned is discussed later; let us first see how they are used. As soon as the bus is available, each device compares its ticket number with the NNS, and the device that finds the match becomes the bus-master. As we explain in the following discussion, there can be only one device whose ticket number matches NNS. Sometime before the bus is available again, NNS is incremented by one. This incrementing is done modulo NTICKETS, so that the ticket numbers range from 0 to (NTICKETS-1). To ensure that devices have distinct ticket numbers, we must have NTICKETS $\geq N$, where N denotes the number of devices.

Now we consider how the ticket numbers are assigned. This is done using NNA. When a device wants to become the bus-master, it copies the NNA into its ticket register. Then, NNA is incremented by one modulo NTICKETS, thus, ensuring that sequentially increasing ticket numbers are "dispensed" in the range from 0 to (NTICKETS-1). Of course, while the copying and incrementing operations are being done, no other device should be allowed to copy the NNA. If the NNA is copied, either there will be two devices with the same ticket number, and confusion will ensue as they both will become bus-masters at a later time, or there will be a device with an invalid ticket number that is outside the above-specified range, and that device would never be able to access the bus, as NNS will never be equal to the invalid ticket number. The accesses to NNA to receive ticket numbers should, therefore, be mutually exclusive.

Thus, in our ticket assignment scheme, achieving mutual exclusion for the bus depends on achieving mutual exclusion at a lower level—that of NNA. The second mutual exclusion is achieved by using one of the existing arbitration schemes; for example, simple daisy-chaining (SDC), rotating daisy-chaining (RDC), modified device polling (MDP), dynamic parallel priority resolution (DPPR), etc. See Ref. 9 for a detailed description and comparison of various bus arbitration schemes.

The duration for which NNA is allocated to a device is the time it takes to copy NNA into its ticket register. This duration is very short—typically, a few gate delays. Thus, the time required to assign a ticket number is essentially the time spent in arbitrating for the use of NNA. Whenever this time is short, as compared to the time for which the main bus is allocated, our scheme would be a close approximation to the ideal FCFS scheme. This is because the devices that request the bus while it is busy are quickly assigned ticket numbers and put into a waiting state. Thus, the scheme remembers the order in which the requests arrive.

It is also possible that some devices will request the bus while NNA arbitration is in progress or while NNA is in use. In such cases, depending upon the NNA arbitration scheme, one among these devices is allowed to copy the next NNA, and they may or may not receive the ticket numbers in the temporal arrival order of their requests. Thus, for the overall scheme to be unbiased, the NNA arbitration scheme must treat the devices in an unbiased way. This is desirable because it is a necessary condition for making the overall scheme a close approximation to the ideal FCFS scheme.

To summarize, the NNA arbitration scheme should be fast and unbiased. We use the criteria in choosing the NNA arbitration scheme. In addition, to judge how close the overall scheme is to the ideal FCFS scheme, we use *s* and *S*. The smaller the value of *s*, the better the approximation, since *s* is minimum in the ideal case. Similarly, the smaller the value of *S*, the better the approximation, since *S* is zero in the ideal case.

We now examine the SDC, RDC, MDP, and DPPR schemes mentioned earlier with regard to their desirability as NNA arbitration schemes.

In SDC, the central arbiter sends out a daisy-chained NNA-grant signal. If a device does not want to access the NNA, it lets the signal pass through; otherwise, it stops the signal and then accesses NNA. Thus, the devices closer to the arbiter are preferred over those farther away from it. This tends to make the overall scheme, named FCFS/SDC, a poorer approximation to the ideal FCFS scheme.

In RDC, the device that accessed NNA last acts as the arbiter for the next arbitration cycle and sends out the NNA-grant signal. On the average, all the devices are given equal treatment, and the average arbitration time is the same as that for SDC. Therefore, the overall scheme, FCFS/RDC, is expected to be a better approximation of the ideal FCFS scheme than the FCFS/SDC.

In MDP, there is no central arbiter, and the daisy-chained NNA-grant signal keeps travelling from device to device in a cyclical fashion. Devices wishing to access the NNA wait for the grant signal to arrive, stop the grant signal temporarily, access the NNA, and then release the grant signal. All the devices receive unbiased treatment. The performance of FCFS/MDP is, therefore, expected to be similar to that of FCFS/RDC, and their hardware implementation is also quite similar.

In DPPR, devices are assigned priorities which change after each NNA arbitration cycle. As the arbitration starts, all the devices that need to access the NNA put their priorities on a common priority bus. Then, each device removes itself from the contention if its priority is lower than the composite priority on the priority bus. This eliminates all but the highest priority device, which then accesses the NNA.[9] The dynamic assignment of priorities in this scheme can be done in a variety of ways, but here we assume it is done so that the order in which devices win arbitration is essentially the same as that of RDC (the same priority assignments emulate MDP also). Initially, the *i*th device is given the priority *i*, and after each arbitration, priorities are cyclically rotated so that the device that won the last arbitration gets the priority one. All the devices are treated equally; however, the average arbitration time for DPPR is much smaller than that of RDC or MDP, because there is no daisy-chained signal involved that gets delayed while passing through each device (by as much as 4 gate delays per device). Thus, FCFS/DPPR is a better approximation to the ideal FCFS scheme than FCFS/RDC and FCFS/MDP. The disadvantage is that FCFS/DPPR requires more hardware than FCFS/RDC or FCFS/MDP.

The conclusions drawn above are supported by the values of s and S obtained through simulation, which are shown in Tables I and II, respectively. (These statistics have been borrowed from Bain and Ahuja.[9]) Both tables include two cases: (*i*) when the schemes discussed above are used for NNA arbitration in the ticket assignment scheme, and (*ii*) when they are used to arbitrate for the main bus itself. Note that the values of s and S for ticket assignment schemes (column 1) are smaller than for the others (column 2); therefore, they are better approximations to the ideal FCFS scheme. Similarly, among the ticket assignment schemes, FCFS/DPPR is the closest approximation to the ideal FCFS scheme. Through simulations, we also observed that as the number of devices is increased, the performance of FCFS/DPPR rapidly converges to that of the ideal FCFS scheme, but the performance of other schemes diverges significantly from that of the ideal FCFS scheme. Therefore, FCFS/DPPR is an attractive scheme when a large number of devices (approximately 16 or more) share a common bus.

III. IMPLEMENTATION OF THE TICKET ASSIGNMENT SCHEME

In this section, we describe and compare two implementations of the ticket assignment scheme. In the first, the NNA and NNS are centralized, and in the second, they are distributed. We consider only the FCFS/MDP scheme, since the implementations with different NNA arbitration schemes are quite similar.

Figure 1a shows an implementation of FCFS/MDP in which the NNA and NNS counters are centralized, and the devices access them through the NNA and NNS buses. If a device requests access to the main bus, it waits for NNA-GT, the cyclically daisy-chained NNA grant signal, to

Table I—Table of s, the standard deviation of the weighting times taken over all requests to the bus. X denotes the schemes in the first column and FCFS/X denotes the ticket assignment using X for NNA arbitration.

X	s for FCFS/X (μs)	s for X (μs)
SDC	19.32	30.26
RDC	1.476	3.214
MDP	1.512	3.230
DPPR	1.368	3.159
Ideal FCFS		1.112

Note: Simulations were carried out for 32 independent devices, each device requesting the bus with uniformly distributed interrequest times between 0.4 and 19.6 μs, with the average interrequest time of 10 μs. The bus was allocated for 0.4 μs for each request.

Table II—Table of S, the standard deviation of the average weighting times experienced by the individual devices. Simulations were carried out under the same conditions as shown in Table I.

X	S for FCFS/X (ns)	S for X (ns)
SDC	524000*	132600†
RDC	68.0	199.9
MDP	63.7	155.8
DPPR	45.0	153.9

* Devices 26 through 32 did not get service.
† Devices 23 through 32 did not get service.

arrive. The device then holds the NNA grant signal, transfers the data on the NNA bus to its ticket register, signals on the INC-NNA line to increment the NNA counter, and then releases the NNA grant signal. (See Fig. 1b for a detailed circuit diagram.) After the device has a ticket number, it waits until the contents of its ticket register are the same as the data on the NNS bus and the bus busy line, BB, is negated. When that occurs, it asserts BB, becomes the bus-master, and signals on the INC-NNS line to increment the NNS counter. After using the bus, it simply negates the BB line. The BB line permits incrementation of NNS to proceed while the main bus is being used.

Figure 2 shows a distributed implementation of the above scheme in which each device has its own NNA and NNS counters. The NNA and NNS buses are eliminated, and the INC-NNA and INC-NNS lines are used to keep the various NNA and NNS counters up to date. The initial values of all the ticket registers, NNA counters, and NNS counters are 0, 1, and 1, respectively.

When a device wants a ticket number, it executes the following sequence of steps:

(*i*) Waits for NNA-GT to arrive, and captures it on arrival.
(*ii*) Initiates steps (*iii*), (*iv*), and (*v*) when INC-NNA becomes false, and does nothing before then.
(*iii*) Shifts the contents of NNA counter into the ticket register.
(*iv*) Signals all the other devices to increment their NNA counters by asserting INC-NNA line. The device also signals itself to do the same. The INC-NNA line is negated after a long enough time to allow the NNA counter to finish incrementing.
(*v*) Releases the NNA-GT signal.

Notice that only steps (*i*) and (*v*) depend on the NNA arbitration

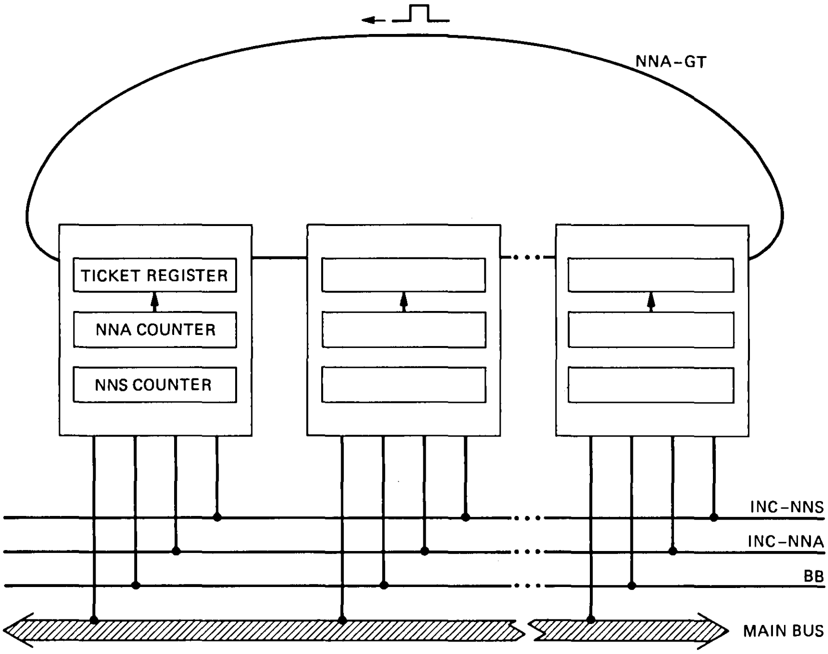


Fig. 1a—An implementation schematic for FCFS/MDP where NNA and NNS are centralized. INC-NNA and INC-NNS are used by the devices to increment NNA and NNS, respectively. NNA-GT is the cyclically daisy-chained grant signal for accessing NNA.

scheme being MDP; they are replaced by a different set of steps for different NNA arbitration schemes. All other steps, including those given below, are independent of the NNA arbitration scheme used. When a device receives the INC-NNA signal, it simply increments the NNA counter.

In the distributed implementation, gaining control of the main bus is similar to that in the centralized implementation:

(*i*) After receiving the ticket number, wait until the contents of the NNS counter and the ticket register are the same, and INC-NNS and BB are false. When that occurs, initiate steps (*ii*) through (*v*); do nothing before then.

(*ii*) Set BB to true.

(*iii*) Signal all other devices to increment their NNS counters by asserting the INC-NNS line. The device also signals itself to do the same. The INC-NNS line is negated after a long enough time to allow the NNS counter to finish incrementing.

(*iv*) Use the main bus.

(*v*) Release the bus by setting BB to false.

As a device receives INC-NN'S, it increments its NNS counter. The detailed circuit diagram for this is similar to Figure 1b, except that each device has NNS and NNA counters of its own.

The distributed implementation has two advantages over the centralized implementation.

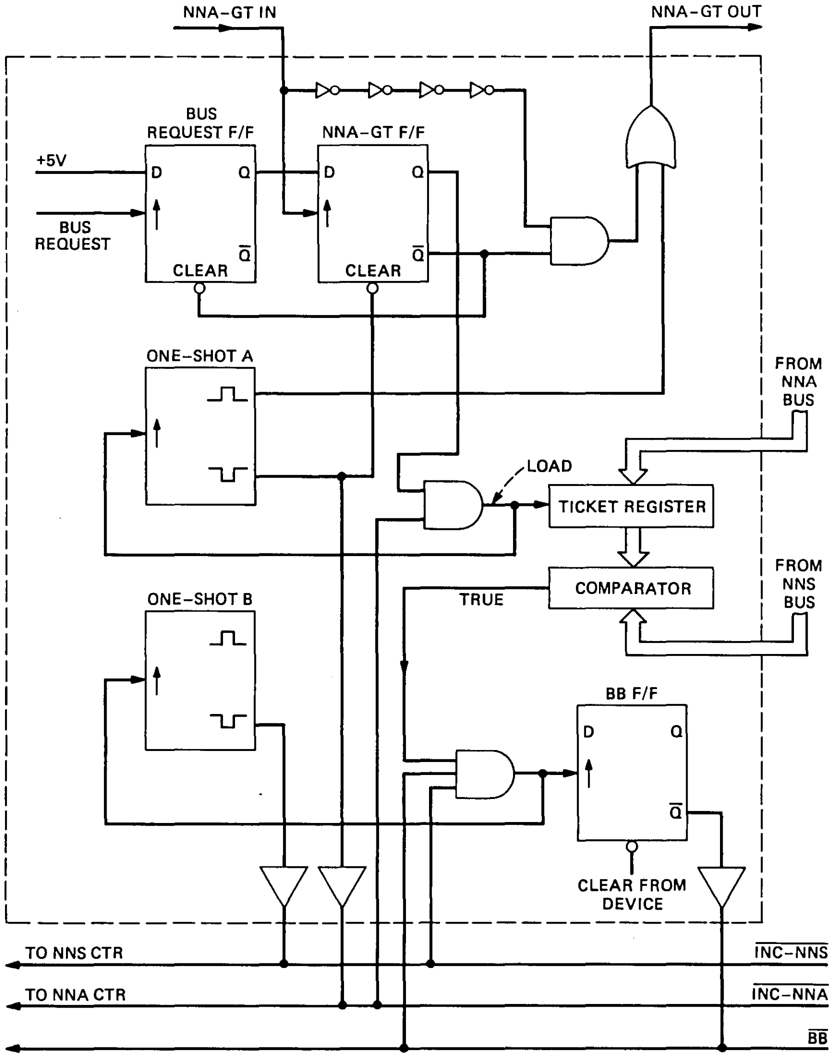


Fig. 1b—The detailed circuit diagram associated with the schematic Fig. 1a. The circuit enclosed within the broken lines is contained in each device. F/F denotes flip-flop. The bus-request F/F and NNA-GT F/F capture the grant signal. The one-shot A generates the outgoing grant signal, the negative of which is also used to signal on the INC-NNA line. The one-shot B generates the INC-NNA signal. Notice that the buses use negative logic.

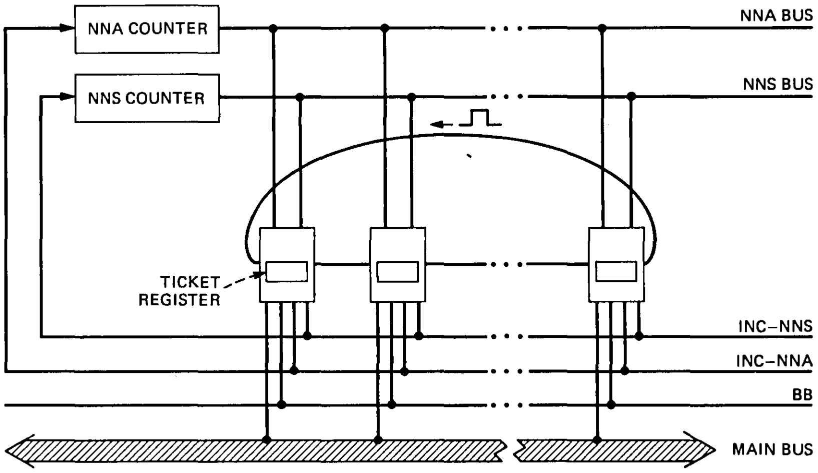


Fig. 2—An implementation schematic for FCFS/MDP where NNA and NNS are distributed. Each device has NNA and NNS registers. INC-NNA and INC-NNS signals are used to keep all the NNA and NNS registers up to date. The NNA and NNS buses have been eliminated.

(*i*) It has fewer lines, since it does not need the NNA and NNS buses.
(*ii*) Devices do not have to wait for voltage levels on the NNA and NNS buses to settle down, as NNA and NNS are available from their local counters. The longer the bus, the more significant this advantage because the settling time of voltages on the bus is proportional to the length of the bus.

The distributed implementation has two disadvantages as compared to the centralized implementation:

(*i*) In order to introduce new devices in the system, their NNA and NNS counters must be current with those in other devices. Although not always satisfactory, this can be done by stopping the system momentarily to reset the counters.
(*ii*) The scheme will malfunction if any one of the counters malfunctions. Depending upon the reliability of the hardware, this disadvantage may not be serious.

Thus, since neither implementation is unequivocally superior to the other, the final choice should be made depending upon the requirements of the application at hand.

IV. SUMMARY

We presented a FCFS bus arbitration scheme that is based upon assigning ticket numbers to the devices as they request the bus. The arbitration for the main bus essentially depends upon the arbitration for the next available ticket number. Several schemes for the latter arbitration were considered, and their impact on the overall scheme was examined using the standard deviation of wait times of all requests and the standard deviation of the average weight times of devices. Using simulation results, we showed that the overall scheme is the closest approximation to the ideal FCFS scheme, when the lower level arbitration is performed by the dynamic, parallel priority-resolution scheme; the resulting overall scheme is called FCFS/DPPR. Two alternative implementations, one centralized and the other distributed, of the overall scheme were described.

V. ACKNOWLEDGMENT

We gratefully acknowledge several fruitful discussions that we had with W. L. Bain, Jr., during the course of this work.

APPENDIX

In the following, we show (*i*) that the ideal FCFS has the minimum value of s, the standard duration of waiting times of all the requests, and (*ii*) that all disciplines of serving the requests have the same, $\bar{w}$, the average waiting time of requests.

Consider the idealized arrangement where the incoming requests are put in a queue in their order of arrival, and the server always picks the first—the oldest—element, in the queue. This is the ideal FCFS scheme. If at any time, the elements in the queue are permuted, we obtain deviations from the ideal case.

Let w_i be the waiting time of the ith request when the queue is not disturbed. Then, for the ideal FCFS scheme,

$$\bar{w} = \frac{1}{N} \sum w_i,$$

and

$$s^2 = \frac{1}{N} \sum (w_i - \bar{w})^2,$$

where N is the total number of requests.

Since any permutation can be expressed as a composition of a number of permutations that exchange two elements, we show that the value of $\bar{w}$ remains the same and that the value of s^2 is increased, if two elements in the queue are interchanged. For simplicity, we assume that the ith and the $(i - 1)$st elements are interchanged. A similar argument holds for the general case also. The new waiting times for these two elements are

$$w_i' = w_i - t,$$

and

$$w_{i-1}' = w_{i-1} + t,$$

where t is the service time for each request. Hence, the difference between the new and the old values of $\bar{w}$ is

$$\begin{aligned}\Delta\bar{w} &= \frac{1}{N}(w_{i-1}' + w_i') - \frac{1}{N}(w_{i-1} + w_i)\\ &= 0.\end{aligned}$$

Also, the difference between the new and the old values of s^2 is

$$\begin{aligned}\Delta s^2 &= \frac{1}{N}[(w_{i-1}' - \bar{w} - \Delta\bar{w})^2 + (w_i' - \bar{w} - \Delta\bar{w})^2]\\ &\quad - \frac{1}{N}[(w_{i-1} - \bar{w})^2 + (w_i - \bar{w})^2]\\ &= \frac{2t}{N}(t + w_{i-1} - w_i).\end{aligned}$$

Note that the maximum value of w_i occurs when the ith request arrives in the queue immediately after the $(i-1)$th request. If the ith request comes later, then the server services some requests in the meantime, thus, reducing the waiting time of the ith request. Hence,

$$w_i \leq w_{i-1} + t.$$

This gives us

$$\Delta s^2 \geq 0,$$

where the equality occurs only when the ith and the $(i-1)$th requests come at the same time. Hence, the ideal FCFS scheme has the minimum value of s^2.

REFERENCES

1. R. C. Chen, "Bus Communication Schemes," NTIS, PB 235-897 (January 1974).
2. K. J. Thurber et al., "A Systematic Approach to the Design of Digital Bussing Structures." AFIPS Conf. Proc. FJCC, 1972.
3. J. V. Levy, "Buses, the Skeletons of Computer Structures," *Computer Engineering*, Bedford, Mass.: Digital Press, 1978, Chapter 11.
4. Digital Equipment Corporation, *VAX-11/780 Architecture Handbook*, 1977.
5. J. W. Conway, "Approach to Unified Bus Architecture Sidestepping, Drawbacks," Comput. Des. *16*, No. 1 (January 1977), pp. 71-6.
6. H. Keller and E. H. Forrester, "Rapid Priority Resolution," U.S. Patent 3,983,540, September 28, 1976.

7. A. G. Fraser, private communication, 1975.
8. K. A. Elmquist, *et al.* “Standard Specifications for S-100 Bus Interface Devices,” *Comput., 12,* No. 7 (July 1979), pp. 28–52.
9. W. L. Bain, Jr., and S. R. Ahuja, “Performance Analysis of Digital Buses for Multiprocessing,” Proc. Eighth Int. Symp. Computer Architecture, Minneapolis, Minn., May 1981.

Copyright © 1981 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 60, No. 7, September 1981
Printed in U.S.A.

Accurate Logic Simulation Models for TTL Totempole and MOS Gates and Tristate Devices

By Y. L. LEVENDEL, P. R. MENON, and C. E. MILLER

(Manuscript received December 16, 1980)

The two logic values, 0, 1, and the unknown, are not sufficient for accurately simulating the behavior of TTL *totempole and* MOS *gates and tristate devices. Furthermore, the classical fault modes (output stuck and input open) are not sufficient to cover the faulty behavior of* MOS *devices. A previous solution to the simulation modeling required the addition of pseudo gates, which have no physical meaning. This paper develops methods of modeling fault-free and faulty tristate devices for logic simulation. The model does not require any additional circuitry, but the existence of a simulator capable of simulating any number of logic values is assumed.*

I. INTRODUCTION

A component finding wide usage in the bus-oriented architecture of today's computer systems is the tristate driver. A typical arrangement is shown in Fig. 1. In this arrangement there are several drivers. Only one driver can be enabled at a time and "talk" to the bus. The receivers capture the information on the bus.

In transistor-transistor logic (TTL) technology, tristate devices allow bus wiring, previously obtained only with conventional open collector output TTL. They also allow the use of active pull up to charge the large capacitances associated with the bus. This feature, not available with conventional bus wiring technique, speeds up the operation of the bus. In MOS technology, similar effects can be achieved with lower power requirements. Here, we shall consider TTL totempole, CMOS, PMOS, and NMOS bistate and tristate devices and show the similarities and differences.

In tristate technology, several malfunctions due to the presence of faults or to a wrong utilization of the bus may occur. These malfunc-

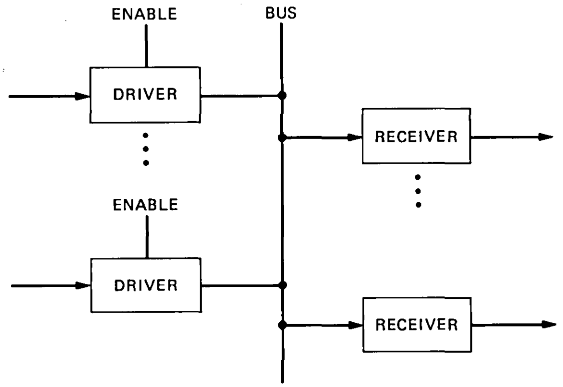


Fig. 1—A tristate bus system.

tions may invalidate test results or damage the components. Therefore, it is important to simulate accurately the operation of faulty and fault-free circuits containing buses, and other tristate devices.

It has been shown that the classical fault modes (output stuck, input open) are not sufficient to cover the faulty and fault-free behavior of CMOS devices.[1,2] One attempt has been made to map these fault effects into classical stuck-type faults by adding circuitry to the fault-free circuit. This additional circuitry is used to provide the faulty and fault-free circuits with memory properties, which exist in CMOS devices under certain conditions. This mapping allows the use of a fault simulator, which simulates only classical faults, for simulating faults in CMOS devices. In fact, the limitations of the available simulator was a constraint on the proposed modeling. Although modeling of the fault-free tristate devices also used similar added circuitry, the effect known as overlap or bus contention which may damage the devices, was not covered by this model.

This paper develops methods of modeling fault-free and faulty tristate devices for logic simulation, without additional circuitry. However, it assumes the existence of a simulator capable of simulating any number of logic values. Both the memory properties of MOS devices and the effects of bus contention are shown to be modeled accurately by the proposed method.

1.1 TTL tristate technology

Consider the tristate inverter of Fig. 2, in which A is the data input and E is the enable lead. The device is enabled and acts as an inverter, when $E = 0$ as shown in Table I. When the inverter is disabled, it assumes a high impedance, namely the impedance between V_o and V_{CC}, and the impedance between V_o and V_G is extremely large. V_o, V_{CC}, and V_G are the output, supply and ground voltages, respectively.

Table I—Truth table for tristate inverter

A	E	V_o
0	0	1
0	1	Disabled
1	0	0
1	1	Disabled

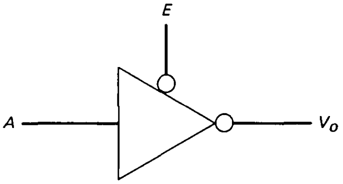


Fig. 2—Tristate inverter.

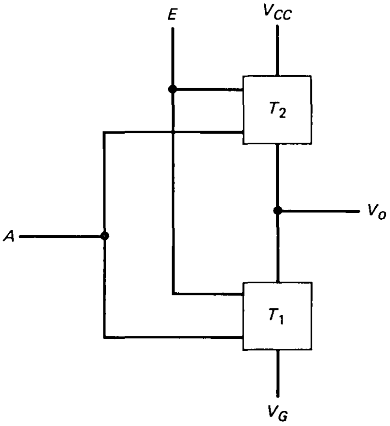


Fig. 3—TTL tristate inverter model.

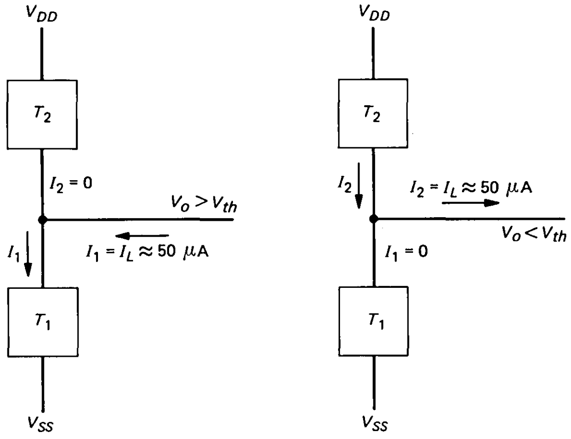
The TTL tristate inverter can be modeled as the connection of two functions T_1 and T_2 (Fig. 3) which are controlled by lines A and E. T_1 and T_2 can be either conducting (on) or nonconducting (off) and they operate according to Table II. The device is in the high-impedance state when both T_1 and T_2 are off. In a tristate bus system, several tristate devices are wired together and the system operates safely if at most one device is enabled at one time (Fig. 1).

Two problems have emerged in tristate bus technology and they are associated with the structure of the tristate devices. The first difficulty concerns a disabled tristate device and its ability to source or sink current depending on the value of its output voltage. These two cases are illustrated in Fig. 4. One can identify a voltage V_{th}, such that, if $V_o > V_{th}$, then T_2 acts as a current source, and if $V_o < V_{th}$, T_1 acts as a sink. Threshold voltage V_{th} is a voltage between V_{CC} and V_G, which is determined by the output properties of the device.

Table II—Truth table for inverter model

A	E	T_1	T_2
0	0	Off	On
0	1	Off	Off
1	0	On	Off
1	1	Off	Off

If a receiver is present on a bus—i.e., it is connected to V_o—the driver will source or sink current and, as a result, the output V_o may reach the input threshold voltage V_{th} of the receiver. The output of a

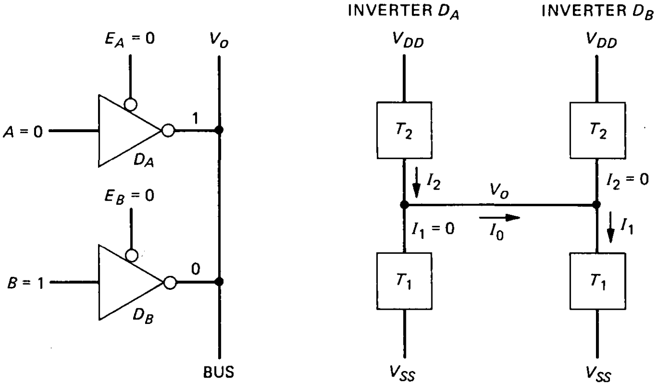


(a) (b)

Fig. 4—Disabled bus. (a) Sink. (b) Source.

receiver with the input voltage equal to V_{th} is unknown and in fact the output may oscillate due to small variations around V_{th}. Normally, after all the driving devices become disabled, the existence of the leakage current I_L will destroy the previous logic value of a bus, and the bus will "float."

A second problem occurs when at least two tristate devices feeding a bus are simultaneously enabled and are in opposite active logic states (Fig. 5). Under this condition, the bus voltage may be anywhere in the range between the active logic levels and the currents may become extremely large (Fig. 5b). The actual value of V_o and I_o can be



(a) (b)

Fig. 5—Bus conflict.

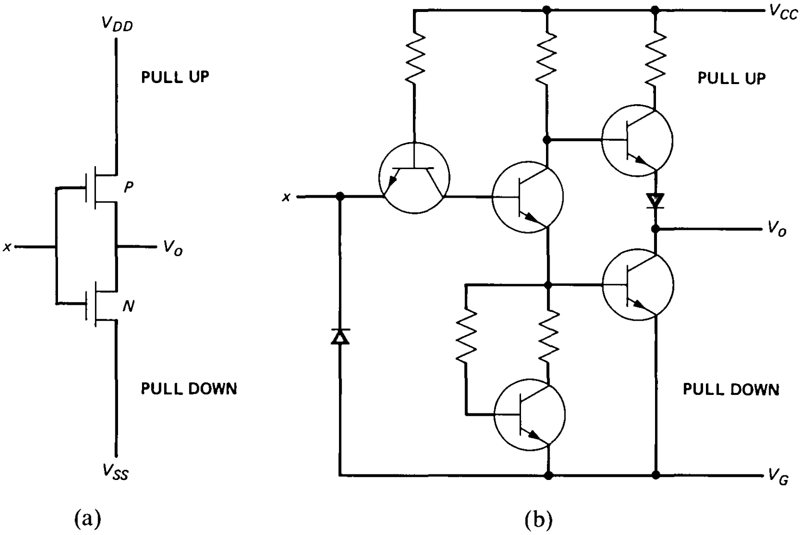
determined from the current-voltage characteristics of the two devices. This condition, called overlap, may cause excessive device heating resulting in device failure or slowly degrade the device, causing a decrease in life.

In simulation, it is important to correctly model the effects of these two problems, so that a simulation user can be warned of the existence of potential difficulties. For instance, a primary input-output bus should be disabled (floating) before a test can be applied to it and enabled before the result of a test can be read from it. Also, an incorrect design or test sequence may cause overlaps on buses and the simulation should produce a warning.

II. TTL TOTEMPOLE, CMOS GATES, AND TRISTATE TRANSMISSION DEVICES

2.1 Pull-up and pull-down functions

Consider the CMOS and TTL implementations of an inverter (Fig. 6). They have a common structure, which can be generalized by the diagram of Fig. 7 for multiple input gates. This structure is composed of a pull-up function (PUF), a pull-down function (PDF), and an integrator (I). The PUF and PDF depend upon the input values $x_1, \cdots, x_n$ and the integrator produces the output Y depending upon Y_U and Y_D. Also the PUF and the PDF are complementary and cannot produce the



(a) (b)

Fig. 6—(a) CMOS inverter. (b) TTL totem pole inverter.

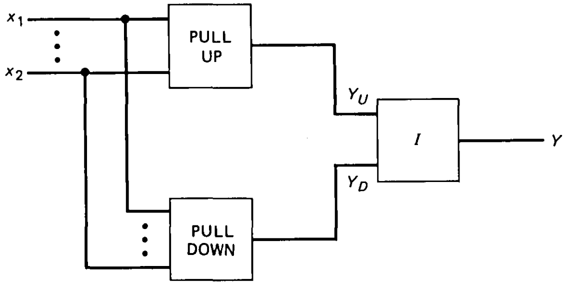


Fig. 7—General model for CMOS and TTL totem pole devices.

same logic value under normal circumstances. By convention, Y_U or Y_D has the value 1(0), when the PUF or the PDF is on(off). The integrator I has the behavior of Table III, except in the case of malfunctions. As an example, consider the CMOS, NAND gate of Fig. 8. The junctions P_1 and P_2 realize a NAND function, whereas N_1 and N_2 realize an AND function.

Table III—Truth table for integrator

Y_U	Y_D	Y
0	0	Impossible
0	1	0
1	0	1
1	1	Impossible

2.2 Tristate devices

A general CMOS or TTL tristate device can be modeled as in Fig. 9. The symbol E represents an enable line, and both PUF and PDF can be simultaneously disabled. Under normal conditions, the PUF and PDF cannot be simultaneously active. The integrator I is described in Table IV.

Table IV—Truth table for tristate integrator

Y_U	Y_D	Y
0	0	High impedance
0	1	Logic 0
1	0	Logic 1
1	1	Impossible

Usually, tristate devices are used in the mode shown in Fig. 10. In the illustration, E_1, E_2, $\cdots$ E_m are the enabling lines. There are several interesting cases, namely

(*i*) All the devices are disabled.

(*ii*) One device is enabled.

(*iii*) Two or more devices are enabled with opposite logic values.

These three cases lead to different impedance situations on the bus and they are represented in Table V. In the context of simulation, it is possible to find a fourth situation, when the impedances are not known. This can be caused by an unknown value on the enable line E of a device.

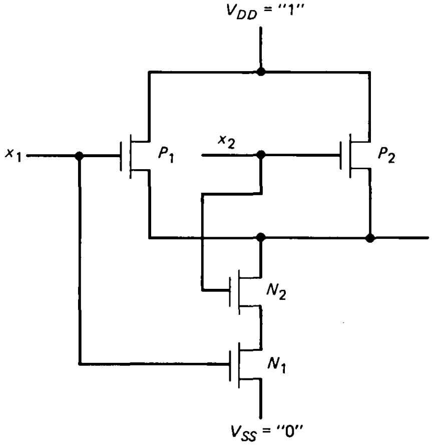


Fig. 8—CMOS NAND gate.

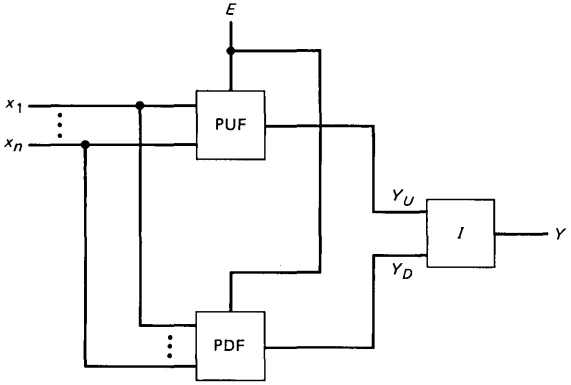


Fig. 9—General model for tristate devices.

Table V—Device states and bus impedances

Devices	Impedance
All disabled	High PUF, high PDF
One enabled	High PUF (PDE), low PDF (PUF)
Two enabled (conflict)	Low PUF, low PDF

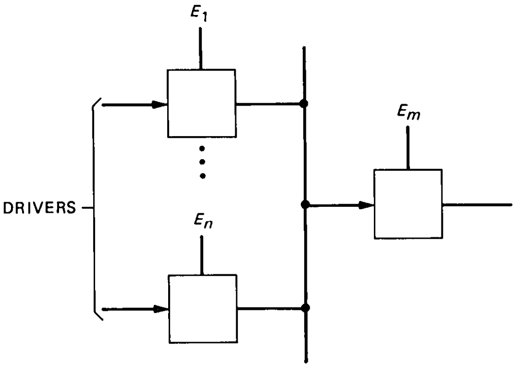


Fig. 10—Tristate devices connected to bus.

2.3 CMOS dynamic properties

The main difference between CMOS and TTL tristate devices is that the leakage currents in CMOS are extremely small compared to TTL leakage currents. Input currents for CMOS are also small. Therefore, if a CMOS device is first enabled and then disabled, the small capacitances on the buses will remain charged for a long period of time and the bus will appear to receivers as if it were remaining at the same logic value. Ultimately, the capacitance will be discharged, but if the rate of operation is sufficiently fast, the discharge time can be considered as infinite and the bus displays memory properties. However, if forced to an active logic value, the bus will immediately reach this value independent of this charged capacitance. In TTL tristate devices, the leakage currents being large, the discharge time becomes small and no memory is displayed.

2.4 Logic values and impedance

It should be clear from our preceding discussion that accurate simulation modeling of tristate devices requires two distinct concepts: logic value and impedance. Three logic values, 0, 1, and u, are widely used in simulation, the symbol u being used to represent unknown signal values.[3] Unknown signal values may be present because the initial values of some leads may be unknown or because of races or oscillations. The effects of impedance on circuit behavior depends on the technology. For instance, a high impedance may appear as an unknown logic value in TTL tristate technology. On the other hand, an output which has a high impedance in CMOS technology will remember the logic value before the gate was disabled. Similarly, a conflict on a bus may appear as a 0, 1, or u depending on the technology.

During simulation of circuits containing tristate devices, it is important to be able to detect special situations like bus conflict. Tests that cause bus conflicts may result in damage to the devices and must be avoided. In the tester environment, the state of an output bus in the high-impedance state may be altered by the tester, invalidating the test. These considerations lead to the representation of the state of a line by a pair composed of the impedance value and the logic value. There will be four possible impedance values (Table VI). Therefore, we obtain 12 combinations of impedance and logic value (Table VII).

Table VI—Impedance values

PUF/PDF	Impedance	Impedance Representation
Both off	High	H
One on, one off	Regular	R
Both on	Conflict	C
One or both unknown	Unknown	U

Table VII—Combinations of impedance and logic values

	Pair	Description
1	R/0	Logic 0
2	R/1	Logic 1
3	R/u	Unknown with a low impedance
4	H/0	High impedance with previous state memory
5	H/1	High impedance with previous state memory
6	H/u	High impedance with unknown previous state memory
7	C/0	Conflict with logic 0 effect
8	C/1	Conflict with logic 1 effect
9	C/u	Conflict with logic u effect
10	U/0	Unknown impedance, logic 0 effect
11	U/1	Unknown impedance, logic 1 effect
12	U/u	Unknown impedance, logic u effect

These 12 logic combinations, which we shall call logic values in the context of simulation, represent a detailed analysis of tristate devices and any one of these corresponds to a possible situation. Two cases are illustrated in Fig. 11 and both cases display memory properties. In the first case (Fig. 11a), the enable line goes from 0 to u and the output goes from R/0 to U/0 (unknown impedance). In the second case (Fig. 11b), the enable line goes to 1 and the impedance goes from R to H, with the same logic value.

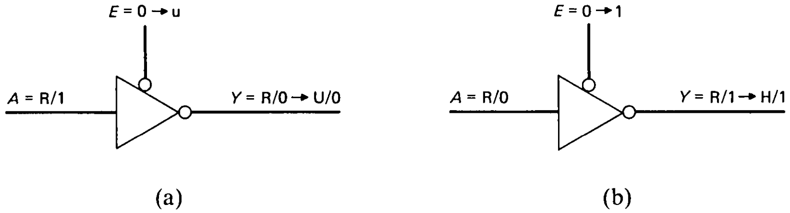


Fig. 11—Determination of impedance.

It is possible to reduce the number of values in Table VII at the expense of some information. The result is Table VIII, which shows two possible sets of logic values for TTL tristate devices. Z and a are synonyms for the pairs H/u and C/u, respectively. In set 2, pairs 3 and 12 are differentiated. Pair 12 is called a potential conflict (a*) and can occur in various situations (Fig. 12). In this case, the simulation will declare a potential bus overlap. In set 1, pairs 3 and 12 are not differentiated and some information may be lost. In TTL technology, the unused combinations correspond to impossible situations (e.g., pair 10) or to unpredictable situations (e.g., pair 8).

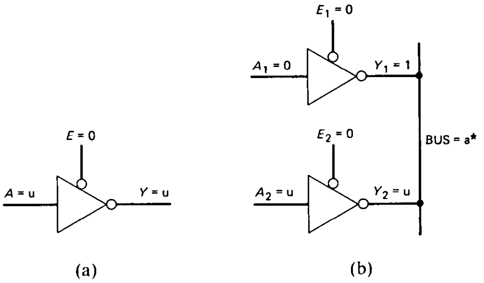


Fig. 12—Potential bus conflict.

Table VIII—Two sets of logic values for TTL tristate devices

Number	Pair	Set 1	Set 2
1	R/0	0	0
2	R/1	1	1
3	R/u	u	u
4	H/0	Unused	Unused
5	H/1	Unused	Unused
6	H/u	Z	Z
7	C/0	Unused	Unused
8	C/1	Unused	Unused
9	C/u	a	a
10	U/0	Unused	Unused
11	U/1	Unused	Unused
12	U/u	u	a*

In the case of CMOS technology, several additional values become meaningful (Table IX). The value Z0 (Z1) is used when a driver having the value 0(1) is disabled and remembers the previous logic value. The value u0(u1) is used when a driver was producing a 0(1) on a bus and the enable line becomes unknown. In all the three sets, pairs 7 and 8 could be used if the actual conflict voltage can be positioned as a 0 or a 1 when added structural knowledge is available.

Table IX—Logic values for CMOS devices

Pair	Set 3
1	0
2	1
3	u
4	Z0
5	Z1
6	Zu or Z
7	Unused
8	Unused
9	a
10	u0
11	u1
12	uu

The sets of values given in Tables VIII and IX can be used instead of impedance/logic value pairs. They are more economical in computer storage and more general; on the other hand, the pair representation may be more efficient, since the impedance is ignored in most of the gate evaluations (except the bus).

We shall illustrate the use of the pairs for a CMOS driver-inverter (Fig. 2). The behavior of the inverter is represented in Table X. *A* and *E* are the input and enable lines, respectively, and *Y* is the output of the device. The symbol *x* represents a "don't care" value.

Table X—Impedance-logic value table for CMOS-driver inverter

Previous Value of *Y*	*AE* = x0	*AE* = 01	*AE* = 11	*AE* = uu
0	H/0	R/1	R/0	U/u
1	H/1	R/1	R/0	U/u
u	H/u	R/1	R/0	U/u

A bus with any number of drivers may be calculated iteratively using Table XI. This table is symmetric with respect to the main diagonal. Its construction is illustrated by the case of three inverters producing *R*/0, *R*/1, and *H*/1, respectively, and wired to a bus. The first pair produces C/u, and C/u is combined with $H/1$ to produce $H/*$, which can be approximated by H/u.

Table XI—Impedance-logic value table for tristate bus

	R/0	R/1	R/u	H/0	H/1	H/u	C/0	C/1	C/u	U/0	U/1	U/u
R/0	R/0	C/*	U/*	R/0	R/0	R/0	C/0	C/*	C/*	U/0	U/*	U/*
R/1	C/*	R/1	U/*	R/1	R/1	R/1	C/*	C/1	C/*	U/*	U/1	U/*
R/u	U/*	U/*	U/u	R/u	R/u	R/u	C/*	C/*	C/u	U/*	U/*	U/u
H/0	R/0	R/1	R/u	H/0	H/*	C/0	C/1	C/u	U/0	U/u	U/u	U/u
H/1	R/0	R/1	R/u	H/*	H/1	H/*	C/*	C/1	C/*	U/u	U/1	C/u
H/u	R/0	R/1	R/u	H/*	H/*	H/u	C/*	C/*	C/u	U/u	U/u	U/u
C/0	C/0	C/*	C/*	C/0	C/*	C/*	C/0	C/*	C/*	C/0	C/*	C/*
C/1	C/*	C/1	C/*	C/1	C/1	C/*	C/*	C/1	C/*	C/*	C/1	C/*
C/u	C/*	C/*	C/u	C/u	C/*	C/u	C/*	C/*	C/u	C/*	C/*	C/u
U/0	U/0	U/*	U/*	U/0	U/u	U/u	C/0	C/*	C/*	U/0	U/u	U/u
U/1	U/*	U/1	U/*	U/u	U/1	U/u	C/*	C/1	C/*	U/u	U/1	U/u
U/u	U/*	U/*	U/u	U/u	U/u	U/u	C/*	C/*	C/u	U/u	U/u	U/u

* Unknown (u) or technology-dependent value.

2.5 Refinement of unknown impedance values

Given that there are three basic impedance values, R, H, and C, the possibility of the enable signal being unknown during simulation introduces indeterminacy in the simulated impedance values. Seven impedance values can be used to represent the three known values and the four cases, where the impedance cannot be uniquely determined. The seven values are:

$$
\begin{aligned}
I_1 &= H\\
I_2 &= R\\
I_3 &= C\\
I_4 &= H \text{ or } R\\
I_5 &= H \text{ or } C\\
I_6 &= R \text{ or } C\\
I_7 &= H \text{ or } R \text{ or } C
\end{aligned}
$$

Figure 13 shows how the indeterminate simulated impedance may be generated. The impedance values are obtained by computing the impedances for a combination of the unknown signal values. For example, in Fig. 13c, it was possible to obtain I_5 because it was known that $E_1 = E_2 = 0$ or 1.

These seven impedance values preserve some information that would otherwise be lost. For instance I_5, I_6, and I_7 represent a potential overlap, whereas I_4 is definitely not an overlap. However, the overhead of dealing with a multiplicity of pairs may not be justified by the gain of information (21 impedance/logic value combinations).

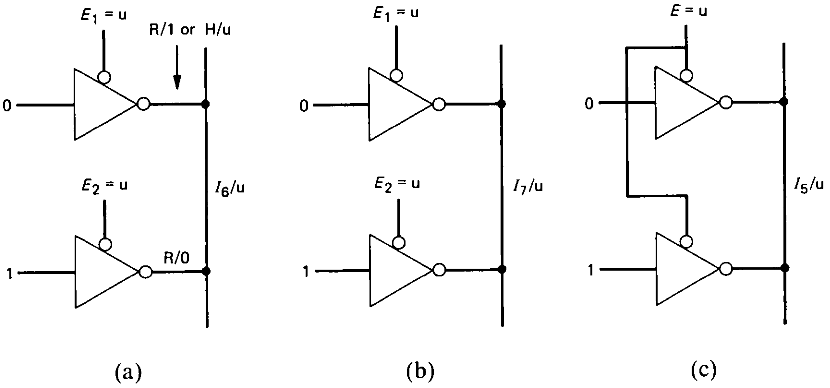


Fig. 13—Generation of indeterminate impedances.

In the above analysis, indeterminacy in the impedance and logic value are treated separately. This also results in some loss of information, which becomes apparent from the computed output of the upper tristate inverter in Fig. 13a. Although its output is known to be $R/1$ or $H/$u, it will be represented by I_4/u. Eliminating this problem would require creating one logic value for each subset of the set: $\{R/0,R/1,R/u,C/0,C/1,C/u,H/0,H/1,H/u\}$, excluding the empty subset. This system would have 2^9—1 or 511 logic values, which is impractical. In Fig. 14, the results would then become $\{C/u,R/0\}$ for case a, $\{C/u,H/u,R/1,R/0\}$ for case b, and $\{C/u,H/u\}$ for case c.

III. FAULT MODELS

Most of the faults that are peculiar to the devices considered in this paper can be simulated using the PUF-PDF model of Figs. 7 or 9.

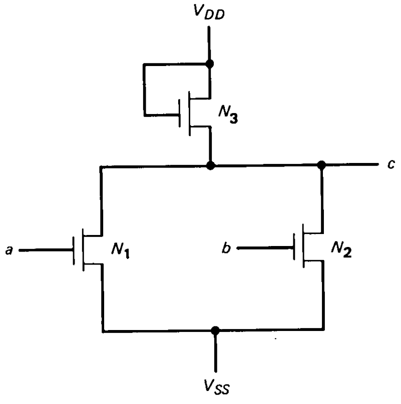


Fig. 14—NMOS NOR gate.

3.1 PUF and PDF faults and enable line faults

We shall consider first a special class of faults, where the PUF or PDF is enabled (disabled) when it is supposed to be disabled (enabled). This class is partitioned into four subclasses described in Table XII. Note that these are single faults.

Table XII—Class of PUF/PDF faults

Fault Subclass	Fault Effect on PUF	Fault Effect on PDF
I	Disabling	No effect
II	No effect	Disabling
III	Enabling	No effect
IV	No effect	Enabling

The fault effects on the output are described in Table XIII. The values marked with * are technology dependent and possibly unknown. Subclasses I and II cause a regular bistate gate to display tristate properties and a tristate device to be disabled when it is supposed to be enabled. Subclasses III and IV may cause a conflict (overlap) under the appropriate input values. In some sense, this fault class blurs the difference between a regular bistate gate and a tristate device. For this reason, we can use the same set of logic values for faulty and fault-free circuit modeling, namely a TTL set (set 1 or 2) or a CMOS set (set 3), for both tristate and bistate devices. The only difference between the two cases is that the high-impedance state will not be produced during the normal operation of a bistate device.

Table XIII—PUF, PDF, and output values

Fault Pull Up or Pull Down	Other Function	Output Value
$Y_U = 0$	$Y_D = 0$	$Y = H/*$
$Y_U = 0$	$Y_D = 1$	$Y = R/0$
$Y_U = 1$	$Y_D = 0$	$Y = R/1$
$Y_U = 1$	$Y_D = 1$	$Y = C/*$
$Y_D = 0$	$Y_U = 0$	$Y = H/*$
$Y_D = 0$	$Y_U = 1$	$Y = R/1$
$Y_D = 1$	$Y_U = 0$	$Y = R/0$
$Y_D = 1$	$Y_U = 1$	$Y = C/*$

Practically, the faults in each subclass can be obtained by shorted or open junctions in the PUF or PDF. We shall consider the example of Fig. 8 and summarize these subclasses of faults in Table XIV. In the context of concurrent fault simulation, the fault-injection mechanism is extremely simple: a fault will be simulated if its effect on the gate is different from the fault-free circuit behavior, and this difference is measured over the set of possible logic values.

Another class of faults can appear in tristate devices and concerns the enable line. An enable line stuck at 0 (stuck at 1) will cause the device to be permanently enabled (disabled). A permanently disabled gate can be modeled by an output "stuck at Z." In the latter case, such a fault is characterized by a Z appearing at an output instead of a known logic value. The faults "enable line stuck at 1" and "output line stuck at Z" are equivalent.

Table XIV—Typical faults in a CMOS NAND gate

Fault Subclass	Fault Example	Input Values	Fault-Free	Fault
I	P_1 open	$x_1 = 0$ $x_2 = 1$	1	Z*
II	N_2 open	$x_1 = 1$ $x_2 = 1$	0	Z*
III	P_2 shorted	$x_1 = 1$ $x_2 = 1$	0	a
IV	N_1 shorted	$x_1 = 0$ $x_2 = 1$	1	a

* With memory of previous logic value.

3.2 PMOS and NMOS devices

The structure of an NMOS (PMOS) device is similar to the CMOS structure in that there is a pull-down function, but the pull-up function is a degenerate case of the CMOS pull-up function, namely it is permanently enabled and serves as a resistor (Fig. 14). The function $c = \overline{a + b}$ is implemented in Fig. 14.

We shall consider several fault modes, namely N_1 open, N_1 shorted, N_3 open, and N_3 shorted. The behavior of these four fault modes is represented in Table XV. The previous and present values of c are denoted by $c(-)$ and c, respectively. The behavior of faults N_1 open, N_1 shorted, and N_3 shorted is independent of $c(-)$. However, if N_3 is open, it is impossible to set c to a one, whereas the combination:

$$\begin{aligned} c(-) &= 0 \\ a &= 0 \\ b &= 0 \end{aligned}$$

produces a disabled output with new logic value equal to previous logic value ($c = Z0$). In spite of the behavioral differences, it is possible to model PMOS and NMOS devices using the same set of logic values as for CMOS devices.

3.3 Input-open faults in CMOS gates

It was shown earlier that a disabled CMOS tristate device displays certain memory properties that could be modeled using additional logic values $Z0$ and $Z1$. When an input to a CMOS gate is open, it is possible to produce these logic values in the faulty circuit. One method of modeling this is by setting the signal value at the site of the fault to a special value "propagating Zi" ($i = 0, 1, u$), and propagating the effect to the gate output. Denoting the propagating $Z0$ and $Z1$ by $PZ0$ and $PZ1$, respectively, we have the following conditions for the generation of these logic values at the site of the input-open fault: when the input changes from 1 or $Z1$ (0 or $Z0$) to 0(1), the faulty value of the input becomes $PZ1$ ($PZ0$). With the introduction of these additional logic values for modeling the fault, we need a method of propagating them through gates. Table XVI shows the NAND function whose inputs are from the set {0, 1, $Z0$, $Z1$, $PZ0$, $PZ1$}. Since we are considering single faults, four entries in Table XVI are undefined.

Table XV—Typical faults in an NMOS NOR gate

			c			
a	b	$c(-)$	Open N_1	Short N_1	Open N_3	Short N_3
0	0	0	1	0	Z0	1
0	0	1	1	0	Impossible	1
0	1	0	0	0	0	a
0	1	1	0	0	0	a
1	0	0	1	0	0	a
1	0	1	1	0	0	a
1	1	0	0	0	0	a
1	1	1	0	0	0	a

This modeling may be applied to the NAND gate of Fig. 8. Consider the fault, junction N_1 open, and x_1 passing from 0 to 1 while $x_2 = 1$. The fault-free output will pass from 1 to 0 and the faulty output will remain at the value $Z1$, meaning that the fault may be detected after the change of x_1 to 1, if it is possible to register $Z1$ as the value 1.

Input-open faults in CMOS gates can also be treated as special types of faults that may produce $Z0$ or $Z1$ on gates outputs depending on gate type and present and previous input values. However, this would require treating input open faults on different types of gates differently. The proposed method presents a uniform way of inserting the effect of the input-open fault and only needs additional logic values during gate evaluation. These logic values introduced for modeling the fault do not themselves reach the gate output.

Table XVI—NAND function with propagating high-impedance states

	0	1	Z0	Z1	PZ0	PZ1
0	1	1	1	1	1	1
1	1	0	1	0	Z1	Z0
Z0	1	1	1	1	1	1
Z1	1	0	1	0	Z1	Z0
PZ0	1	Z1	1	Z1	—	—
PZ1	1	Z0	1	Z0	—	—

IV. CONCLUSIONS

A general model consisting of a pull-up function, a pull-down function and an integrator is proposed for modeling TTL totempole, CMOS, PMOS, and NMOS devices. It is shown that an accurate representation of the state of tristate devices and also certain bistate devices require not only the logic values but also impedance values. A set of 12 combinations of impedances and logic values is proposed, each of which can be represented by a single value or by an impedance/logic value pair. Speed-storage trade-offs will determine the choice of representation. The set of logic values needed is shown to be a technology-dependent subset of the 12 combinations represented. The proposed model covers all the known tristate fault-free and faulty effects and does not require any additional modeling gates.

REFERENCES

1. R. L. Wadsack, "Fault Modeling and Logic Simulation of CMOS and MOS Integrated Circuits," B.S.T.J., *57,* No. 5 (May-June 1978) pp. 1449-74.
2. R. H. Krambeck, private communication.
3. J. S. Jephson, R. P. McQuarrie, and R. E. Vogelsberg, "A Three-Value Computer Design Verification System," IBM System J., *8,* No. 3 (1969), pp. 178-188.
4. E. B. Eichelberger, "Hazard Detection in Combinational and Sequential Switching Circuits, IBM J. Res. Develop., *9,* No. 2 (March 1965), pp. 90-9.
5. E. G. Ulrich and T. Baker, "The Concurrent Simulation of Nearly Identical Digital Networks," Computer, *7,* No. 4 (April 1974), pp. 39-44.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 60,, No. 7, September 1981
Printed in U.S.A.

Special-Information-Tone Frequency Detection

By A. FEUER

(Manuscript received January 20, 1981)

The ability to distinguish recorded announcements from other call types is an essential part of the mechanized service-evaluation process. To do this, all announcements will have a special-information-tone prefix. A frequency detector—which is based on the correlation functions of the received signal—would be used to decide which announcement was triggered by a specific call attempt. This paper evaluates the performance of the frequency detector in the presence of additive noise and frequency shift induced by the announcement machine. The theoretical results, based on a calibration frequency, are very encouraging. To verify that use of this frequency is feasible in practice, an algorithm is proposed and its performance evaluated to show that it compares favorably to the theoretical one.

I. INTRODUCTION

As part of the process of evaluating the end-to-end performance in the telephone network, a sample of call attempts is evaluated and the attempts are classified into several categories, such as completed, busy, recorded announcements, etc. To mechanize this classification process, a machine must have the ability to distinguish between completed calls and recorded announcements. Current planning for the mechanized systems envisions the use of special-information-tone (SIT) prefixes which are to be attached to recorded announcements and can then be automatically recognized by the mechanized classifier (as well as alert the customer to the fact that he is listening to a recorded announcement). By choosing four distinct SITs, each representing a certain category of recorded announcements, the classifier will have the ability to distinguish between these categories as well as to recognize a recorded announcement in general.

The SIT is defined as a sequence of three consecutive tones. To get four distinct SITs five frequencies were chosen: $\underline{f}_1 < \underline{f}_2 < \overline{f}_1 < \overline{f}_2 < f_3$, and each SIT consists of $\underline{f}_i$, $\overline{f}_j$, f_3. The third tone has a fixed frequency and will be used for calibration as is described later. The actual values of these frequencies are 904.5 Hz, 985.4 Hz, 1356.8 Hz, 1440.2 Hz, and 1758.5 Hz. (The choice of these frequencies is mainly the result of constraints imposed by the CCITT definition of special information tones and tone generation and detection considerations.) To recognize which of the possible SITs was received, the machine has to detect in the first tone whether it was $\underline{f_1}$ or $\underline{f_2}$ and in the second tone whether it was $\overline{f_1}$ or $\overline{f_2}$. This means that the classification process of the announcement categories is reduced to the two frequency-detection processes mentioned above. However, the planned direct recording of the SIT followed by the recorded announcement on the various announcement machines in the telephone network may introduce significant degradation into the reproduced SIT. In addition to additive noise, which is common to all signals in the network, frequency flutter and frequency shift of considerable effect on the reproduced tones may occur. The flutter effect, having an oscillatory nature, can be minimized by averaging properly the received data. In this paper, we report on our investigation of the combined effects of additive noise and frequency shift on the detection process. The detection scheme considered here involves use of correlation functions, and we have concluded that a reliable detection of SIT frequencies is possible provided that a certain level of signal-to-noise ratio is ensured and the available data is properly used. It should be pointed out that in order to carry out the analysis we assume that the additive noise is white. We recognize, however, that in reality a colored additive noise may have a dominating effect on the error bounds.

The structure of the paper is as follows. Section II presents the basic ideas of frequency detection via the correlation functions. In Section III, a white additive noise is considered to be present and the performance of the detector in this case as a function of signal-to-noise ratio is evaluated. The effects of frequency shift are introduced in Section IV and the use of the calibration frequency to eliminate these effects is considered. The performance of the frequency detector as a function of the signal-to-noise ratio is evaluated and the results of using the calibration frequency are compared to the results when it is not used; a significant improvement is observed.

The performance evaluations in Sections III and IV are based on using the likelihood-ratio test. This, because of the complexity of the expressions involved, is not practically feasible; however, these evaluations provide bounds on the performance of other schemes. In Section V, an algorithm for a detection process is presented. This algorithm is simple enough to be practical and its performance compares favorably to the theoretical one.

II. CORRELATION DETECTOR

The use of correlation functions of a received signal for detection purposes was devised as part of the overall classification process by J. E. Walls of Bell Laboratories.[1] Some of the principles of the classification process that are relevant to our analysis are briefly described here.

Let $r(t)$ be a received signal observed for a period of length T. The correlation functions are defined by

$$C_n = \frac{1}{T}\int_{T/2}^{T/2} r(t)r(t + n\Delta)dt \qquad n = 0, 1, 2 \cdots, \tag{1}$$

where $\Delta < T$ (practically we want Δ to be small relative to T).

We note that C_0 is in fact the signal's average power in this time interval.

Since we are interested in frequency detection, let us assume for a moment that the received signal is a pure frequency, namely

$$r(t) = \sin 2\pi ft$$

and derive expressions for the correlation functions in this case. Then,

$$C_n = \frac{1}{T}\int_{T/2}^{T/2} \sin 2\pi ft \cdot \sin 2\pi f(t + n\Delta)dt,$$

or carrying out the simple integration results in

$$C_n = \frac{\cos 2\pi fn\Delta}{2}\left[1 - \frac{\sin 4\pi fT}{4\pi fT}\right]. \tag{2}$$

The power term C_0 and the next two terms C_1, C_2, are used for the detection process. Since

$$C_0 = \frac{1}{2}\left[1 - \frac{\sin 4\pi ft}{4\pi ft}\right]$$

$$C_1 = C_0 \cos 2\pi f\Delta$$

$$C_2 = C_0 \cos 2(2\pi f\Delta),$$

if we normalize C_1 and C_2 by the power C_0 we get the following relationship

$$\left[\frac{C_2}{C_0}\right] = 2\left[\frac{C_1}{C_0}\right]^2 - 1. \tag{3}$$

Expression (3) means that all pure frequencies correspond to a point on a parabola in the $(C_1/C_0, C_2/C_0)$ plane. Clearly this parabola exists only for $-1 \leq C_1/C_0 \leq 1$ and $-1 \leqq C_2/C_0 \leq 1$ (see Fig. 1) and C_1/C_0,

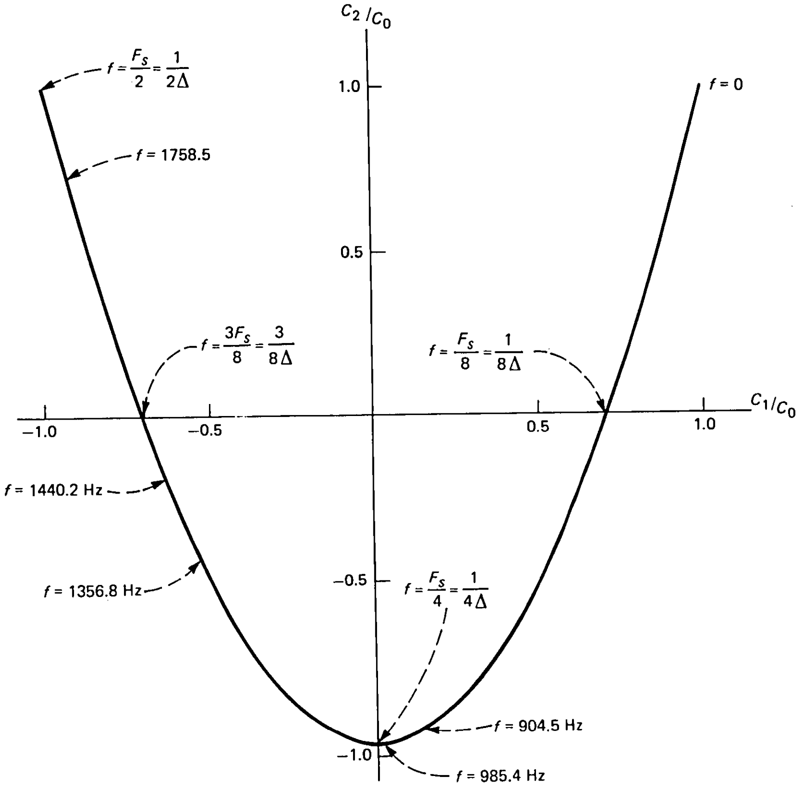


Fig. 1—Correspondence between various frequencies and the parabola on the (C_1/C_0, C_2/C_0) plane.

C_2/C_0 are periodic functions of the frequency, so more than one frequency may correspond to the same point on the parabola. This means that for the detection process one must be concerned with the uniqueness of the points corresponding to the frequencies involved. For instance, if the two frequencies involved are $f_1 = \frac{1}{4}\,\Delta$ and $f_2 = \frac{3}{4}\,\Delta$, both correspond to the point (0, −1) in the (C_1/C_0, C_2/C_0) plane. This means that computing C_0, C_1, C_2 in this case is not sufficient to make the detection between these two frequencies possible even in the deterministic case (absence of noise).

However, the choice of the parameter $1/\Delta$ for a given set of frequencies involved in the detection process can ensure the necessary uniqueness. One way of doing it is to choose $1/\Delta$ to be larger than twice the largest frequency to be detected, and this will be sufficient, as can be seen in Fig. 1.

In the presence of noise, as is shown later, the point will shift from the parabola towards the origin, while the origin itself corresponds to white noise for which both C_1 and C_2 are equal to zero.

In practice, instead of integration, the received signal is sampled and the sampled values are then used to get a good approximation of the correlation functions. Let $r(t)$ again denote the received signal, sampled with frequency $F_s = 1/\Delta$, and N the number of samples done in the observation interval T. Then eq. (1) will be replaced by

$$C_n = \frac{1}{N} \sum_{i=0}^{N-n-1} r_i r_{i+n}, \tag{4}$$

where

$$r_i = r\left[\frac{i}{F_s}\right].$$

Again, for a pure frequency, after some manipulations it can be shown that

$$C_n = \frac{1}{N} \sum_{i=0}^{N-n-1} \sin 2\pi \frac{f}{F_s} i \sin 2\pi \frac{f}{F_s} (i+n)$$

can be written as

$$C_n = \frac{1}{2} \cos n\, 2\pi \frac{f}{F_s} \left[\frac{N-n}{N} - \frac{1}{N} \cos(N-1) 2\pi \frac{f}{F_s} \frac{\sin N\, 2\pi \frac{f}{F_s}}{\sin 2\pi \frac{f}{F_s}}\right]$$

$$+ \frac{1}{2} N \cos(N-1) 2\pi \frac{f}{F_s} \frac{\cos N\, 2\pi \frac{f}{F_s} \sin n\, 2\pi \frac{f}{F_s}}{\sin 2\pi \frac{f}{F_s}}. \tag{5}$$

With the assumption that N is large compared to $n = 0, 1, 2$, we may write

$$C_n = \frac{1}{2} \cos 2\pi \frac{f}{F_s} n \left[1 - \frac{1}{N} \cos(N-1) 2\pi \frac{f}{F_s} \frac{\sin N\, 2\pi \frac{f}{F_s}}{\sin 2\pi \frac{f}{F_s}}\right], \tag{6}$$

so that the relationship eq. (3) holds here as well. In the Appendix, the values of the correlation functions computed by eqs. (1) and (6) are given for the frequencies used for SIT. The error values introduced by going from eqs. (5) to (6) are also given in the Appendix for the same frequencies.

Suppose now that a frequency, one of possible two, is transmitted.

If on the receiving end we make sure that the sampling frequency is large enough compared to the two possible frequencies, we ensure the uniqueness of the correspondence between these frequencies and the parabola defined by eq. (3). With this assurance, we can sample the received signal, compute C_0, C_1, C_2 according to eq. (4), and then use the values C_1/C_0 and C_2/C_0 to decide which frequency was initially sent.

III. ADDITIVE NOISE

In the previous section, frequency detection using correlation functions was described. As long as the received signal contains nothing but the pure frequency originally sent, the problem is clearly deterministic. However, in practice there are various degradations that affect the received signal. In this section, we are interested in evaluating the performance of the correlation detector in the presence of additive white noise.

Thus, let the received signal be

$$r(t) = A \sin 2\pi f_k t + n(t) \qquad k = 1, 2,$$

where A corresponds to the signal-to-noise ratio and $n(t)$ is a normalized white Gaussian noise. The sampled values are

$$r_i = A \sin \theta_k\, i + n_i, \tag{7}$$

where

$$r_i = r\left[\frac{i}{F_s}\right],\ n_i = n\left[\frac{i}{F_s}\right],\ \theta_k = 2\pi\frac{f_k}{F_s},$$

$$k = 1, 2, \qquad i = 0, 1, 2, \cdots, N-1,$$

and n_i are independent identically distributed (IID) variables with zero mean and variance one.

Substitution of eq. (7) into eq. (4) results in

$$C_n^k = \frac{1}{N}\left[\sum_{i=0}^{N-n-1} A^2 \sin i\theta_k \sin(i+n)\theta_k + \sum_{i=0}^{N-n-1} A[n_{i+n}\sin i\theta + n_i \sin(i+n)\theta] + \sum_{i=0}^{N-n-1} n_i n_{i+n}\right]. \tag{8}$$

Since eq. (8) contains random variables, C_0, C_1, C_2, and hence, the ratios C_1/C_0 and C_2/C_0 are also random variables with certain density functions. Once these density functions are known the detection process becomes a standard problem described in detail in various textbooks on detection theory (see for example Ref. 2). We use the likelihood ratio test. To compute the threshold, we assume that the frequencies to be detected have equal a priori probabilities. The costs are assumed to be zero and one for correct detection and error, respectively. With the knowledge of the density functions and the threshold, we can evaluate the performance of the detection process by computing the error probability.

The first step then, is to develop the expressions for the density functions of C_1/C_0 and C_2/C_0. However, in view of the complexity of eq. (8), rather than attempt to develop an exact expression, we will make use of a version of the Central Limit Theorem (Theorem 4.2.5 in Ref. 3) and develop approximate expressions for these density functions.

Let us now denote

$$S_n^k = \frac{A^2}{N} \sum_{i=0}^{N-n-1} \sin i\theta_k \sin(i+n)\theta_k. \tag{9a}$$

$$Y_{n1}^k = \frac{A}{N} \sum_{i=0}^{N-n-1} [n_{i+n} \sin i\theta_k + n_i \sin(i+n)\theta_k]. \tag{9b}$$

$$Y_{n2} = \frac{1}{N} \sum_{i=0}^{N-n-1} n_i n_{i+n}. \tag{9c}$$

Then,

$$C_n^k = S_n^k + Y_{n1}^k + Y_{n2}. \tag{10}$$

With some manipulation, we can get more convenient expressions for S_n^k and Y_{n1}^k:

$$S_n^k = \frac{A^2}{2} \cos n\theta_k \left[\frac{N-n}{N} - \frac{1}{N} \cos(N-1)\theta_k \frac{\sin N\theta_k}{\sin \theta_k} \right] + \frac{A^2}{2} N \cos(N-1)\theta_k \frac{\sin n\theta_k}{\sin \theta_k}.$$

$$Y_{n1}^k = 2\frac{A}{N} \cos n\theta_k \sum_{i=n}^{N-n-1} n_i \sin i\theta_k + \frac{A}{N} \left[\sum_{i=0}^{n-1} n_i \sin(i+n)\theta_k + \sum_{N-n}^{N-1} n_i \sin(i-n)\theta_k \right].$$

Now, since $N >> n$ we can write

$$S_n^k = \frac{A^2}{2} \cos n\theta_k \left[1 - \frac{1}{N} \cos(N-1)\theta_k \frac{\sin N\theta_k}{\sin \theta_k} \right]$$

and

$$Y_{n1}^k = 2\frac{A}{N} \cos n\theta_k \sum_{i=0}^{N-1} n_i \sin i\theta_k,$$

or as

$$S_0^k = \frac{A^2}{2}\left[1 - \frac{1}{N}\cos(N-1)\theta_k \frac{\sin N\theta_k}{\sin\theta_k}\right] \tag{11a}$$

$$Y_{01}^k = 2\frac{A}{N}\sum_{i=0}^{N-1} n_i \sin i\theta_k. \tag{11b}$$

We get

$$S_n^k = S_0^k \cos n\theta_k \tag{12a}$$

$$Y_{n1}^k = Y_{01}^k \cos n\theta_k. \tag{12b}$$

Denoting

$$x_1^k = \frac{C_1^k}{C_0^k}$$

$$x_2^k = \frac{C_2^k}{C_0^k}$$

and using eqs. (10) and (12) we get

$$x_1^k = \frac{[S_0^k + Y_{01}^k]\cos\theta_k + Y_{12}^k}{S_0^k + Y_{01}^k + Y_{02}^k} \tag{13a}$$

$$x_2^k = \frac{[S_0^k + Y_{01}^k]\cos 2\theta_k + Y_{22}^k}{S_0^k + Y_{01}^k + Y_{02}^k}. \tag{13b}$$

We observe now that x_1^k and x_2^k are functions of the random variables Y_{01}^k, Y_{02}^k, Y_{12}^k, Y_{22}^k, each one of which is a linear combination of IID random variables. The commonly used version of the Central Limit Theorem can be applied to conclude that the above four variables have Gaussian limiting distributions. This in turn enables us to use Theorem 4.2.5 in Ref. 3 to find the distributions of x_1^k and x_2^k.

The first step is to compute the means and variances of Y_{01}^k, Y_{02}^k, Y_{12}^k, and Y_{22}^k. Using the fact that the n_i's are IID Gaussian random variables with zero mean and variance one and using eqs. (9c) and (11b) we can readily verify that if

$$Y^k = \begin{bmatrix} Y_{01}^k \\ Y_{02}^k \\ Y_{12}^k \\ Y_{22}^k \end{bmatrix},$$

then

$$E\{Y^k\} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

and

$$\operatorname{cov}(Y^k) = E\left[[Y^k - E\{Y^k\}][Y^k - E\{Y^k\}]' \right] = \begin{bmatrix} \frac{4}{N} S_0^k & 0 & 0 & 0 \\ 0 & \frac{2}{N} & 0 & 0 \\ 0 & 0 & \frac{1}{N} & 0 \\ 0 & 0 & 0 & \frac{1}{N} \end{bmatrix}. \quad (14)$$

As in Ref. 3 denote vector b:

$$b = E[Y^k] = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}.$$

From eq. (14), the entries of Y^k, being Gaussian independent random variables, are jointly Gaussian, as is $\sqrt{N}(Y^k - b)$ with zero mean and covariance matrix

$$T = N \begin{bmatrix} \frac{4}{N} S_0^k & 0 & 0 & 0 \\ 0 & \frac{2}{N} & 0 & 0 \\ 0 & 0 & \frac{1}{N} & 0 \\ 0 & 0 & 0 & \frac{1}{N} \end{bmatrix}.$$

Now, let

$$x^k(Y^k) = \begin{bmatrix} x_1^k \\ x_2^k \end{bmatrix} \quad (15)$$

and

$$\phi_{ij} = \frac{1}{\sqrt{N}} \frac{\partial x_i^k}{\partial Y_j^k} \bigg| b,$$

so by eq. (13)

$$\phi = \frac{1}{\sqrt{N}[S_0^k + 1]^2} \begin{bmatrix} \cos\theta_k & \cos 2\theta_k \\ -S_0^k \cos\theta_k & -S_0^k \cos 2\theta_k \\ S_0^k + 1 & 0 \\ 0 & S_0^k + 1 \end{bmatrix}.$$

All this establishes the preconditions for Theorem 4.2.5 in Ref. 3, which then states that the vector $[x^k(Y^k) - x^k(b)]$ has a Gaussian limiting distribution with zero mean and covariance matrix

$$R = \phi' T \phi. \tag{16}$$

Since N in our case is quite large, we may write

$$p_{x^k}(x^k) = p_{x|f_k}(x|f_k) = \frac{1}{2\pi|R_k|^{1/2}} \exp\left[-\frac{1}{2}[x - x^k(b)]' R_k^{-1} [x - x^k(b)] \right], \tag{17}$$

where

$$R_k = \frac{1}{N(S_0^k + 1)^4} \begin{bmatrix} (S_0^k)^2[2\cos^2\theta_k + 1] + 2S_0^k[2\cos^2\theta_k + 1] + 1, & 2S_0^k\cos\theta_k\cos 2\theta_k(S_0^k + 2), \\ 2S_0^k\cos\theta_k\cos 2\theta_k(S_0^k + 2), & (S_0^k)^2[2\cos^2 2\theta_k + 1] + 2S_0^k[2\cos^2 2\theta_k + 1] + 1 \end{bmatrix}$$

and

$$E\{x^k\} = x^k(b) = \frac{S_0^k}{S_0^k + 1} \begin{bmatrix} \cos\theta_k \\ \cos 2\theta_k \end{bmatrix}.$$

From eqs. (11a) and (17), we can readily observe the way the signal-to-noise ratio, A, affects the density function. If A gets very large, S_0^k becomes very large; then the mean value of x^k approaches the parabola and the entries of R_k become very small—namely, we approach the deterministic case described in the previous section. On the other hand, if A approaches zero, S_0^k goes to zero as well and the mean value approaches the origin. This is understandable since as A gets smaller, the effect of the noise gets larger, dominating the signal; and, being a white noise, the cross correlation functions C_1 and C_2 approach zero.

In Figs. 2 and 3, we see the effect of A as described above for two pairs of frequencies that were selected for the SIT. The sampling frequency is 4000 Hz. It should be noted that the mean values for each frequency are on a straight line as is obvious from eq. (17). In the figures, for every A, one equal-probability contour is drawn around the mean and the fact that these contours get smaller as A gets larger is expected since the entries of R_k are getting smaller as was pointed out earlier.

Once we have eq. (17) the detection is straightforward. With a priori probabilities and costs as described earlier, the measured data is used to compute the point in the $(C_1/C_0, C_2/C_0)$ plane and test which conditional density has higher value at this point. Then decide on the corresponding frequency as the one that was sent. More details of this procedure are described in Ref. 2. However, using the expressions we have for the density functions the process can be somewhat simplified. From eq. (17) and Ref. 2,

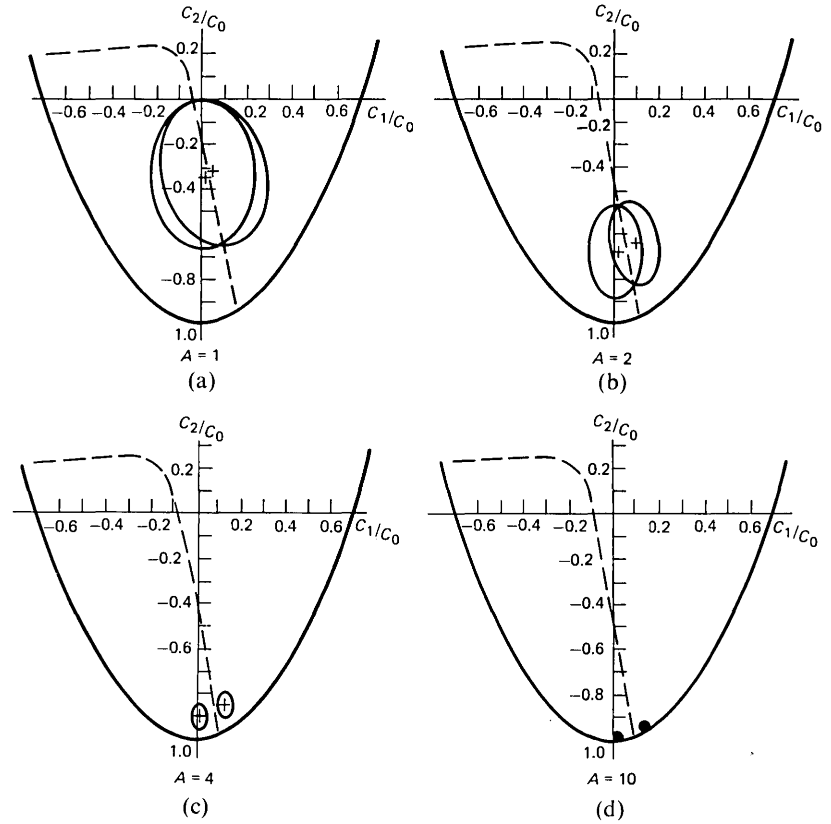


Fig. 2—Dependence of the density functions $P(x\,|\,f)$ on signal-to-noise ratio for 904.5 and 985.4 Hz (dashed lines are the equal likelihood points).

$$p_{x\,|\,f_1}(x\,|\,f_1) \underset{f_2}{\overset{f_1}{\gtrless}} p_{x\,|\,f_2}(x\,|\,f_2),$$

which means that if the inequality holds in one direction, we decide on f_1, and if in the other, f_2. This is equivalent to

$$\mathrm{1n}|R_2| - \mathrm{1n}|R_1| + [x - x^2(b)]'R_2^{-1}[x - x^2(b)] - [x - x^1(b)]'R_1^{-1}[x - x^1(b)] \underset{f_2}{\overset{f_1}{\gtrless}} 0, \qquad (18)$$

which with an equality sign is a line in the $(C_1/C_0,\ C_2/C_0)$ plane. The dotted lines in Figs. 2 and 3 correspond to the above-mentioned line. On one side of these lines one density function has higher values, on the other side the other density function. The detection then consists of deciding on which side of this line the measured point falls.

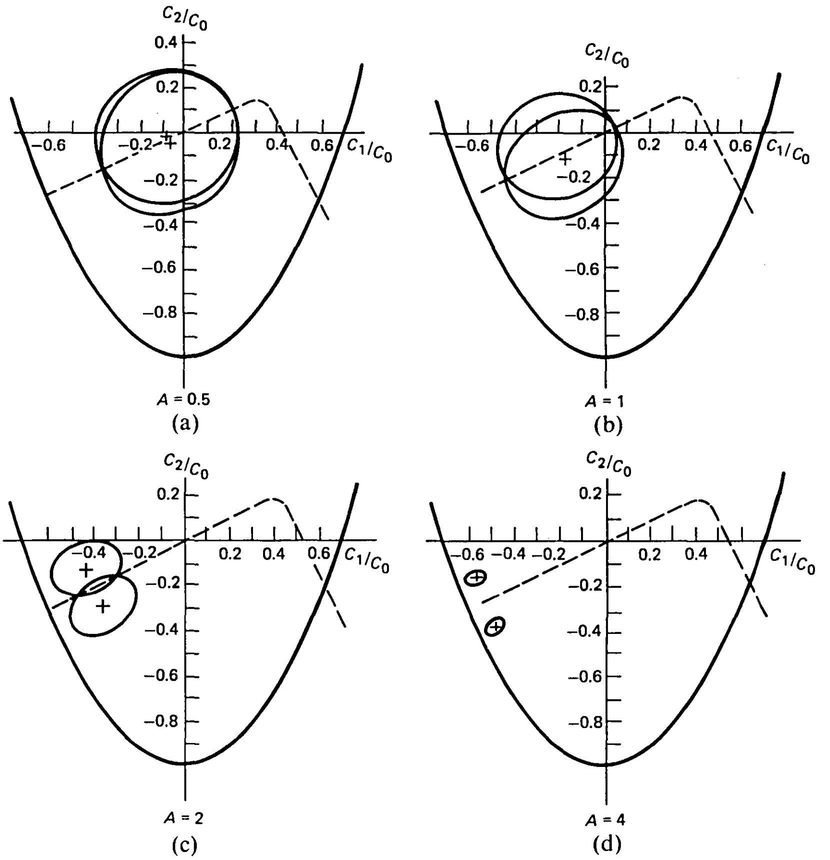


Fig. 3—Dependence of the density functions $p(x \mid f)$ on signal-to-noise ratio for 1356.8 and 1440.2 Hz (dashed lines are the equal likelihood points).

To evaluate the performance of this detector, we compute the error probability. Let B_k, $k = 1, 2$ be the part of the plane in which $p_{x \mid f_k}(x \mid f_k)$ has the larger value. Then the probability of error will be

$$\Pr(\epsilon) = \frac{1}{2}\left[\int_{B_2} p_{x \mid f_1}(x \mid f_1)dx + \int_{B_1} p_{x \mid f_2}(x \mid f_2)dx\right]. \tag{19}$$

Figure 4 shows the error probability as a function of A for, again, the two pairs of frequencies of interest in the SITs 904.5 Hz, 985.4 Hz, and 1356.8 Hz, 1440.2 Hz, and sampling frequency of 4000 Hz. Since, in general, signals transferred by the telephone network result in signal-to-noise ratios higher than 8 dB, the results here are encouraging for both frequency pairs. It is interesting to note that the higher frequency pair results in a better performance. Observing Figs. 2 and 3, one can see the reason for this difference in performance; the density functions for every signal-to-noise ratio are better separated in the higher pair. It turns out, that for every choice of a sampling frequency and difference in value, some pairs of frequencies—and not necessarily the lower valued frequencies—are more detectable than others.

IV. ADDITIVE NOISE AND FREQUENCY SHIFT

The use of announcement machines as tone generators will introduce two primary types of degradation, frequency flutter and frequency shift. In this section, we attempt to analyze the effects of the frequency shift with additive noise on the performance of the correlation detector. It is assumed that the flutter effect is eliminated by averaging a number of observations of each received tone.

The complexity of the expressions involved in the analysis here makes closed-form results very difficult if possible at all. For this reason, digital computer calculations were used as the main tool in the analysis.

Introducing the frequency shift, we get the expression for the received signal

$$r(t) = A \sin 2\pi \frac{(1 + n_d) f}{F_s} t + n(t),$$

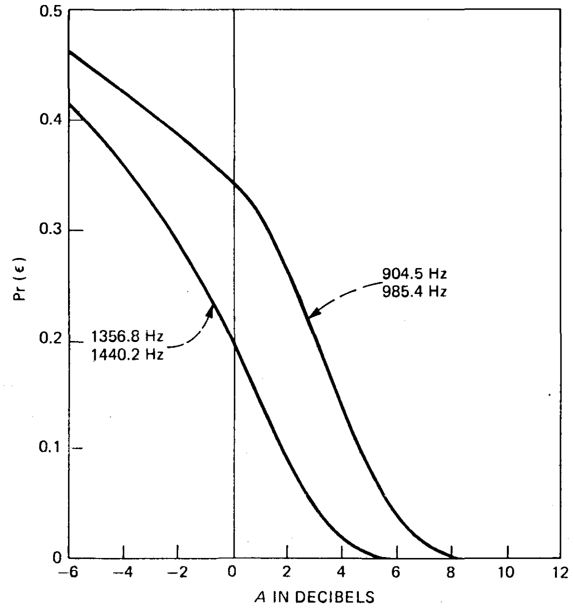


Fig. 4—Probability of error when no frequency shift is considered.

where n_d is assumed to be a random variable with extended beta distribution. This particular distribution is general enough to include many possibilities and agrees with the physical properties of n_d (namely $-1 \leq n_d \leq 1$). So n_d's assumed density function is

$$p_{n_d}(n_d) = \begin{cases} B(1+n_d)^{\alpha-1}(1-n_d)^{\beta-1} & \text{for} \quad |n_d| \leq 1 \\ 0 & \text{elsewhere} \end{cases},$$

where

$$B = \left|\frac{1}{2}\right|^{\alpha+\beta+1} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

and $\alpha,\beta \geq 1$ are the two distribution parameters (in our computations we chose $\alpha = \beta = 10$ which fits reasonably the little data available for n_d).

The presence of this additional degradation causes the expressions developed in the previous section to be conditional on n_d. This means that now, rather than having an expression for $p_{x|f_k}(\cdot|\cdot)$, we have an expression for $p_{x|n_d,f_k}(\cdot|\cdot,\cdot)$ or using eq. (17) we may write

$$p_{x|n_d,f_k}(x|n_d,f_k) = \frac{1}{2\pi|R_k|^{1/2}} \exp\left[-\frac{1}{2}[x - E\{x^k\}]'R_k^{-1}[x - E\{x^k\}]\right], \quad (20)$$

where the expressions for R_k, $E\{x^k\}$, S_0^k are as before [see (11a) and (17)] and $\theta_k = 2\pi \frac{(1+n_d)f_k}{F_s}$.

Since for the likelihood ratio test $p_{x|f_k}(\cdot|\cdot)$ is needed, we can proceed to compute it using the relation

$$p_{x|f_k}(x|f_k) = \int_{-1}^{1} p_{x|n_d,f_k}(x|n_d,f_k)p_{n_d}(n_d)dn_d.$$

With this and eq. (19), the performance of the detector can be evaluated for this case where both additive noise and frequency shift are present. In Fig. 5 the probabilities of error in detection as a function of the signal-to-noise ratio are presented. Comparing these results to Fig. 4 makes it clear that the performance of the detector deteriorates very significantly when frequency shift is present. Even for a very high signal-to-noise ratio the frequency shift induces considerable error. The effect of the frequency shift alone, which provides lower bounds on the error probabilities in Fig. 5, can be calculated as follows. When A gets very large the received signal becomes a pure tone with a shifted frequency. This, however, implies that in this case it will be sufficient to consider only $x_1 = C_1/C_0$ for the detection. Since x_1 and n_d are

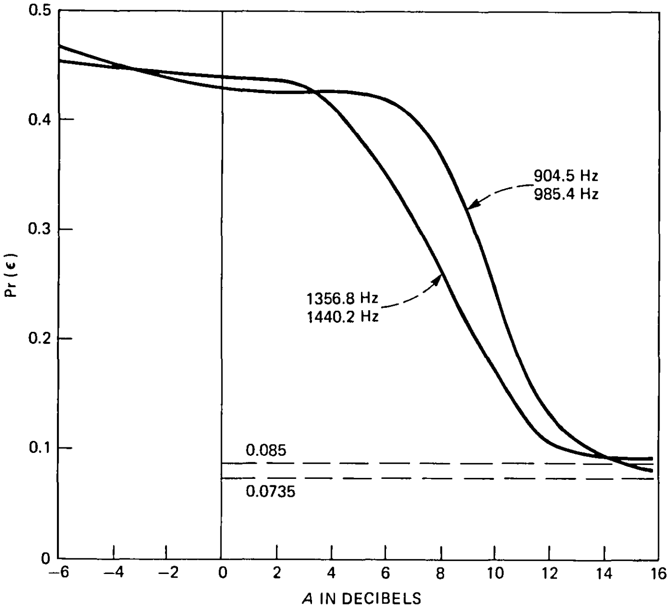


Fig. 5—Probability of error when frequency shift is considered but no calibration frequency is used.

related through

$$x_1 = \cos 2\pi(1 + n_d)\frac{f_k}{F_s},$$

the knowledge of $p_{n_d}(\cdot)$ makes the calculation of $p_{x_1|f_{ki}}(\cdot\,|\,\cdot)$ for each frequency f_k, straightforward. In Fig. 6 the density functions for the two pairs of frequencies are drawn and the thresholds for the detection in this case are the intersections of the density functions (also pointed out in the figure).

The technique to overcome the frequency shift is based on using the third tone for calibration in the detection of the first two tones. This means that a fixed tone is sent, processed in the receiving end, and the knowledge of its exact value and the corresponding measurements can be used to improve the detection of the first two tones.

To be more precise, let f_3 be the frequency of the third tone and $x^3 = [C_1^3/C_0^3,\, C_2^3/C_0^3]$ computed from the corresponding measured data. Then through eq. (20) we know that

$$p_{x^3|n_d,f_3}(x^3\,|\,n_d, f_3) = \frac{1}{2\pi|R_3|^{1/2}}$$

$$\exp\left[-\frac{1}{2}[x^3 - E\{x^3\}]'R_3^{-1}[x^3 - E\{x^3\}]\right],$$

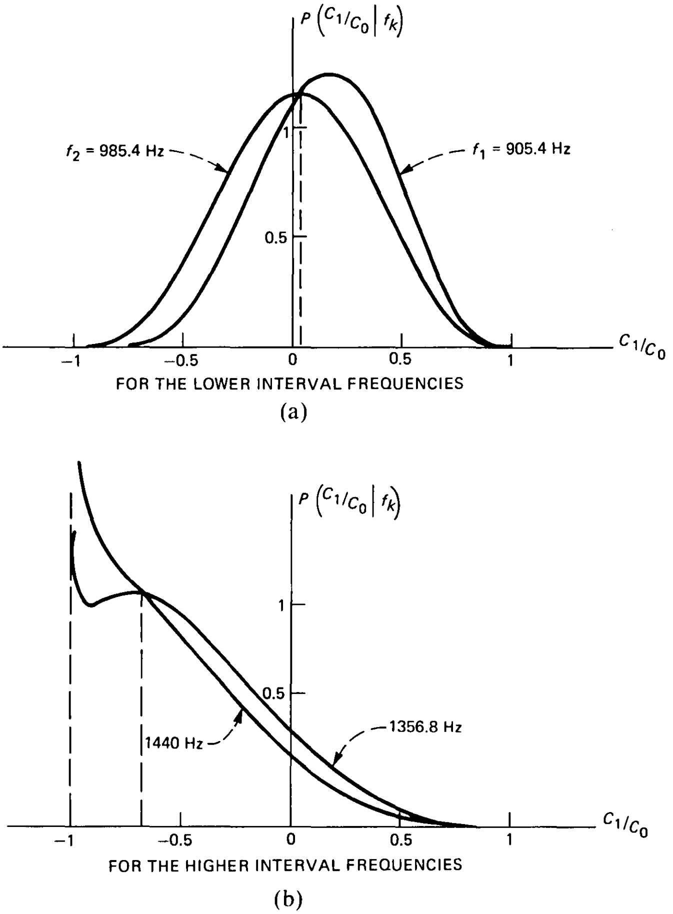


Fig. 6—Density functions required for detection when frequency shift is considered but the additive noise is ignored ($A \rightarrow \infty$).

where the expressions for $E\{x^3\}$ and R_3 are as in eqs. (11a) and (17) with f_3 substituted for f_k. Using this we can improve on our knowledge of the statistics of n_d by computing $p_{n_d|x^3,f_3}(.\,|\,.,.)$ through the relationship

$$p_{n_d|x^3,f_3}[n_d|x^3, f_3] = \frac{p_{x^3|n_d,f_3}(x^3|n_d, f_3)p_{n_d}(n_d)}{\displaystyle\int_{-1}^{1} p_{x^3|n_d,f_3}[x^3|n_d, f_3]p_{n_d}(n_d)dn_d}.$$

This improved data on n_d can then be used to calculate $p_{x|x^3,f_3,f_k}$ $(.\,|\,.,.,.)$

$$p_{x|x^3,f_3,f_k}[x|x^3, f_k] = \int_{-1}^{1} p_{x|n_d,f_k}(x|n_d, f_k)p_{n_d|x^3,f_3}[n_d|x^3, f_3]dn_d,$$

which in turn will be used for the detection process, namely

$$p_{x|x^3,f_3,f_1}[x|x^3, f_3, f_1] \underset{f_2}{\overset{f_1}{\gtrless}} p_{x|x^3,f_3,f_2}[x|x^3, f_3, f_2]. \tag{21}$$

In Figs. 7 and 8, we present—for this improved detection process with 1758.5 Hz as the calibration frequency—the error probabilities as a function of the signal-to-noise ratio. Note that the sampling frequency 4000 Hz is larger than 2 × 1758.5. Comparing this to the results without the calibration frequency, we observe considerable improvement, whereas the results computed with no frequency shift present provide a lower bound on error probabilities (or an upper bound on the improved detector's performance).

The case when $A \to \infty$ (i.e., when the additive noise becomes negligible) is again of special interest but very simple. Since now with the knowledge of both x_1^3 and f_3 the shift can be exactly computed,

$$x_1^3 = \cos 2\pi(1 + n_d)\frac{f_3}{F_s}$$

or

$$n_d = \frac{F_s}{2\pi f_3}\cos^{-1}x_1^3 - 1,$$

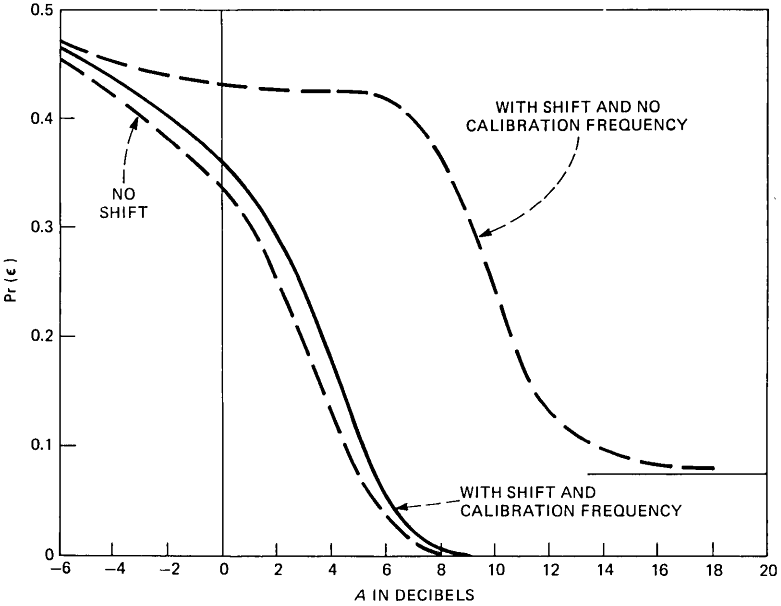


Fig. 7—Error probabilities with frequency shift and calibration frequency—lower interval frequencies.

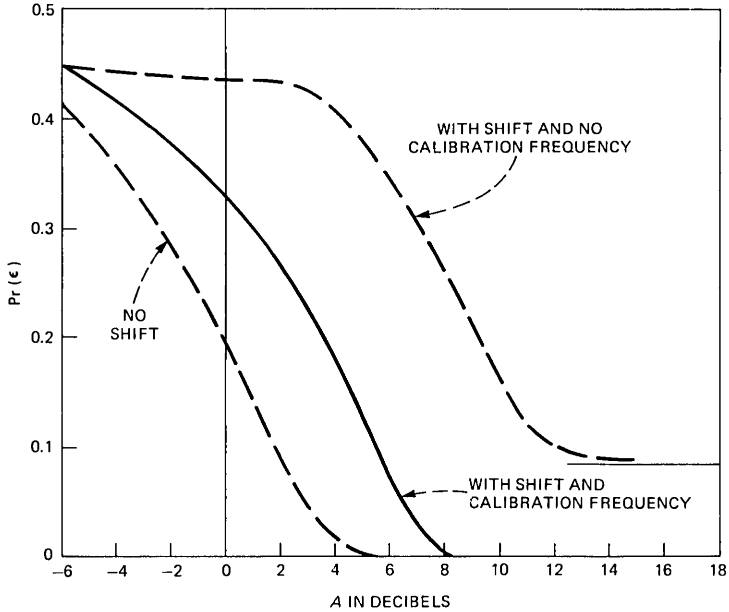


Fig. 8—Error probabilities with frequency shift and calibration frequency—higher interval frequencies.

and the detection of the first two tones becomes deterministic. The only source of problems in this case is the question of uniqueness of $\cos^{-1}x_1^3$. This can be overcome by considering both $\cos^{-1}x_1^3$ and $2\pi - \cos^{-1}x_1^3$ and with probability one, one of them will recover one of the two possible frequencies.

V. A SUGGESTED PRACTICAL CALIBRATION ALGORITHM

In the previous section, we described how the use of a calibration frequency can theoretically improve the detection of tones that are affected by frequency shift and additive noise. However, practical computation of the density functions required in eq. (21), or the separating line (where the two density functions are equal) is impossible. Hence, an algorithm is required that is both compatible to the theoretical exact density function separation and simple enough to be practically implemented. Such an algorithm is presented and its performance evaluated.

Let us first review the available data to be used for the detection. The received sequence of three tones is sampled and for each tone the correlation functions are calculated repeatedly and averaged to eliminate frequency flutter effect. Then the cross-correlation functions are normalized to give the three vectors

$$x^1 = \begin{bmatrix} \dfrac{C_1^1}{C_0^1} \\ \\ \dfrac{C_2^1}{C_0^1} \end{bmatrix} \quad x^2 = \begin{bmatrix} \dfrac{C_1^2}{C_0^2} \\ \\ \dfrac{C_2^2}{C_0^2} \end{bmatrix} \quad x^3 = \begin{bmatrix} \dfrac{C_1^3}{C_0^3} \\ \\ \dfrac{C_2^3}{C_0^3} \end{bmatrix}$$

corresponding to the three tones. Whereas the frequencies resulting in x^1 and x^2, respectively, are not known (in each case they can be either one of two possible frequencies), the third one is known to result from 1758.8 Hz. The idea is to use this knowledge to get an estimate of the frequency shift to help in the decision process of the first two tones.

The suggested algorithm (see Fig. 9) is as follows:

(*i*) Draw a line from the origin through x^3; it will intersect the parabola at

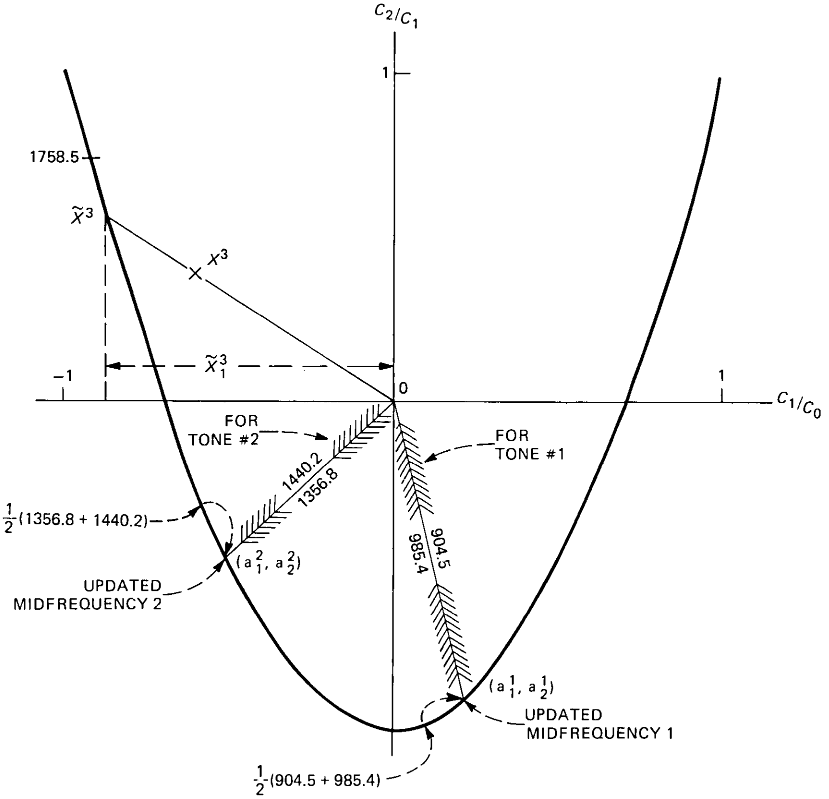


Fig. 9—Geometric interpretation of suggested algorithm.

$$\tilde{x}^3 = \begin{bmatrix} \tilde{x}_1^3 \\ \\ \tilde{x}_2^3 \end{bmatrix} = \begin{bmatrix} 1 \\ \\ \dfrac{x_2^3}{x_1^3} \end{bmatrix} \tilde{x}_1^3,$$

where

$$\tilde{x}_1^3 = \frac{x_2^3 + \sqrt{[x_2^3]^2 + 8[x_1^3]^2}}{2x_1^3}. \tag{22}$$

(*ii*) Use $\tilde{x}_1^3$ to compute the values

$$\left.\begin{aligned} a_1^i &= \cos \theta_i \\ a_2^i &= \cos 2\theta_i \end{aligned}\right\} i = 1, 2, \tag{23}$$

where

$$\begin{aligned} \theta_1 &= \frac{904.5 + 985.4}{2 \times 1758.5} \cos^{-1} \tilde{x}_1^3 \\ &= 0.5374 \cos^{-1} \tilde{x}_1^3 \end{aligned} \tag{24a}$$

and

$$\begin{aligned} \theta_2 &= \frac{1356.8 + 1440.2}{2 \times 1758.5} \cos^{-1} \tilde{x}_1^3 \\ &= 0.7953 \cos^{-1} \tilde{x}_1^3. \end{aligned} \tag{24b}$$

(*iii*) Draw the lines from the origin to the points (a_1^i, a_2^i): If x^i is counterclockwise away from the corresponding line, the *i*th tone is of the lower frequency, and if it is clockwise away, it is of the higher frequency.

The motivation behind this algorithm is quite simple. We have observed earlier that the additive noise effect is to shift the mean value of the pair $(C_1/C_0, C_2/C_0)$ towards the origin, whereas the frequency shift causes this pair to move along the parabola. The first step in the algorithm can be viewed then as isolation of the effect of the frequency shift. The point $\tilde{x}^3$ on the parabola is regarded as the result of the original frequency 1758.5 Hz, together with some shift that can now be estimated using the relationship

$$\tilde{x}_1^3 = \cos 2\pi \frac{1758.5}{4000} (1 + \hat{n}_d)$$

or

$$\hat{n}_d = \frac{4000}{2\pi\, 1758.5} \cos^{-1} \tilde{x}_1^3 - 1.$$

This estimate is then used to update the midfrequencies for each tone, ½(904.5 + 985.4)(1 + n_d) for the first and ½(1356.8 + 1440.2)(1 + n_d) for the second. The lines that connect the origin with the points

$$\begin{bmatrix} \cos\dfrac{\pi(904.5 + 985.4)}{4000}(1 + \hat{n}_d) \\ \cos\dfrac{2\pi(904.5 + 985.4)}{4000}(1 + \hat{n}_d) \end{bmatrix}$$

and

$$\begin{bmatrix} \cos\dfrac{\pi(1356.8 + 1440.2)}{4000}(1 + \hat{n}_d) \\ \cos\dfrac{2\pi(1356.8 + 1440.2)}{4000}(1 + \hat{n}_d) \end{bmatrix}$$

provide the threshold lines for the detection of the frequencies of the respective tones. In Fig. 9 the algorithm is described geometrically.

In Figs. 10 and 11 the performance of a detector using this algorithm is compared to the performance when exact separation is assumed. It is quite clear that the use of the algorithm results in a performance that is very close to the theoretical one.

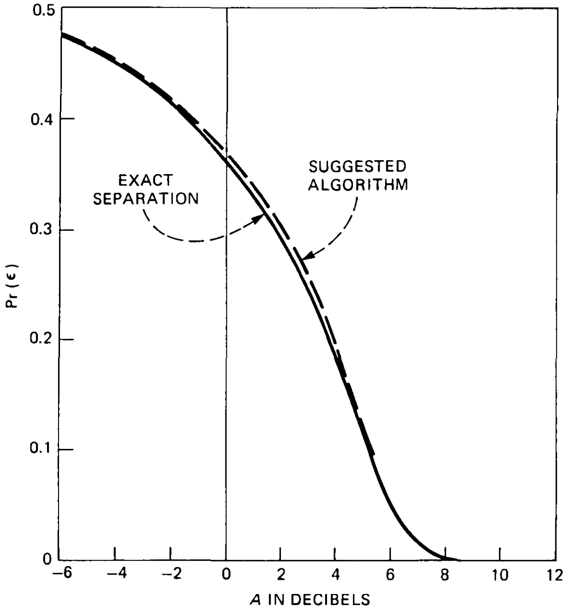


Fig. 10—Comparison of error probabilities for suggested detection algorithm to the exact separation—lower interval frequencies.

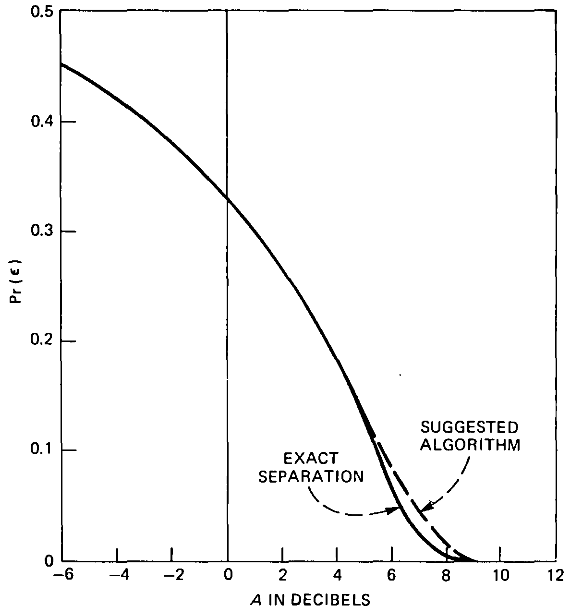


Fig. 11—Comparison of error probabilities for suggested detection algorithm to the exact separation—higher interval frequencies.

There are two difficulties in the described algorithm that will affect its performance. The first is because positive shifts above approximately 14 percent will result in frequencies higher than the critical one—$F_s/2$ (see Appendix)—for the third tone, and the points on the parabola are no longer uniquely related to their corresponding frequencies. This difficulty is inherent to the correlation function approach and can be eliminated only if a significantly higher sampling frequency is chosen (approximately 7000 Hz). However, if we assume that the shift is always less than the critical one, even for the proposed sampling frequency, the effect does not seem to be significant since in the model we have chosen for the frequency shift the probability of having shifts higher than 14 percent is quite low. The second difficulty, which is inherent in the proposed algorithm, arises when the resultant $\tilde{x}_1^3$ is less than -1. In this case, we propose simply to take it equal to -1 and again argue that the probability of this happening is very low even for small signal-to-noise ratios. Altogether, both difficulties, if treated as is suggested above, do not seem to affect the performance of the detection algorithm.

VI. CONCLUSION

In this paper, we addressed some of the problems in detecting recorded announcements encoded via the special information tones

(SIT). In particular, we discussed problems that arise when additive noise is present. We have assumed that the frequency flutter effects are eliminated by averaging several observations, and investigated in detail only the additive noise and frequency shift effects.

Our results support the conclusion that by properly using the information on the frequency shift, its effects can be made almost negligible and under these conditions high-performance SIT detection can be achieved.

The performance evaluations presented here make explicit use of a certain assumed model for the noise and frequency shift; however, the detection algorithm, which is proposed in Section V, is independent of any such assumptions. The performance of this algorithm is comparable to that theoretically achievable, and thus this algorithm is proposed for implementation in the SIT classification process.

VII. ACKNOWLEDGMENTS

The author gratefully acknowledges Mr. Chuck Antoniak for his useful comments and observations.

APPENDIX

Approximations Used in Computing the Correlation Functions by Sampled Data.

f (Hz)	θ (rad)	CC_0	CC_1	CC_2	C_0	C_1	C_2	E_1	E_2
904.5	1.42	0.5	0.0747	−0.4778	0.4971	0.0743	−0.4749	0.0001	0.00004
985.4	1.55	0.499	0.0114	−0.4985	0.4982	0.0114	−0.4976	0.0008	0.00003
1356.8	2.13	0.4993	−0.2654	−0.2171	0.499	−0.2653	−0.217	0.0008	−0.0008
1440.2	2.26	0.4993	−0.3184	−0.0932	0.4997	−0.3186	−0.0933	0.0003	−0.0004
1758.5	2.76	0.5004	−0.4649	−0.3632	0.496	−0.4608	0.36	−0.0019	0.0034

Continuously computed correlation functions [see eq. (2)]:

$$CC_n = \frac{1}{2}\cos(n\theta)\left[1 - \frac{\sin 4\pi ft}{4\pi ft}\right]$$

Approximated correlation functions [see eq. (6)]:

$$C_n = \frac{1}{2}\cos(n\theta)\left\{1 - \frac{1}{N}\cos[(N-1)\theta]\frac{\sin(N\theta)}{\sin\theta}\right\}$$

Differences between discretely computed correlation functions [see eq. (5)] and their approximated values.

$$E_n = \frac{1}{2N}\cos[(N-1)\theta]\frac{\sin(n\theta)}{\sin\theta}\cos(N\theta).$$

[Note: $F_s = 4000$ Hz; $N = 167$; $T = 1/24$ seconds and $\theta = 2\pi\,(f/F_s)$].

REFERENCES

1. J. E. Walls, private communication.
2. L. H. Van Trees, *Detection, Estimation, and Modulation Theory*, Part 1, New York: John Wiley, 1968.
3. T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, Canada: John Wiley, 1958.

Copyright © 1981 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 60, No. 7, September 1981
Printed in U.S.A.

Effect of Echo Canceler on Common-Channel Interoffice Signaling Continuity Check

By G. S. FANG

(Manuscript received December 12, 1980)

Annoying echoes generated by impedance mismatches can occur in long-distance telephone networks. An echo canceler controls echo by synthesizing an echo replica and subtracting it from the actual echo on the return path. For the echo canceler to work properly, it must be able to distinguish between desired signals and echoes. One instance where this can be difficult is during the common-channel interoffice signaling (CCIS) continuity check between four-wire switching offices. On CCIS trunks the voice and the signaling are routed separately. To ensure a satisfactory transmission path, a voice path continuity check is conducted before call setup. The check is performed on a loop basis by sending a check tone and looping it back at the distant office. Since the echo canceler adaptive filter memorizes information from the last previous call, it may partially cancel the looped-back tone. In this paper we study the effect of the echo canceler adaptive memory on the CCIS continuity check. The analytical and experimental results indicate that the occurrence of continuity check failure caused by the presence of an active echo canceler is so infrequent and insignificant compared to the existing statistics of all other pertinent failure mechanisms. Thus, disabling the echo canceler during the CCIS continuity check is unnecessary.

I. INTRODUCTION

An echo may be produced in a transmission system whenever there is an impedance mismatch. This impedance discontinuity can cause a significant portion of the transmitted signal energy to be reflected toward the signal source over an echo path. Noticeable echoes constitute one of the most serious forms of impairment in telephone channels. Its subjective annoyance increases with both echo amplitude and propagation delay. The increasing use of satellite circuits for domestic and international calls has made the control of echoes even more important. The traditional approach of echo control using via net loss or echo suppressors is not acceptable for full-hop satellite circuits, i.e., circuits carrying both directions of transmission via satellite.[1,2] A better approach is to use echo cancelers that control echo by synthesizing a replica of the echo and subtracting it from the returned signal.[2,3] The realization of the single-chip integrated echo canceler has made its widespread deployment appealing.[4]

Figure 1 models the essential elements of the echo path to show how echo cancelers work. When there is far-end speech $x(t)$ but no near-end speech $v(t)$, the internal registers of the echo canceler adaptively update the estimate of the echo path impulse response $h(t)$ to form an echo replica $\hat{y}(t)$ that is subtracted from the real echo $y(t)$. When the near-end speech is detected in the presence of the far-end speech (double talk) the speech detector inhibits further updating, but the echo canceler still tries to cancel the echo contained in $y(t)$ by using the most recent estimate of the echo path impulse response. This property of continued echo canceling during double talk is a nice feature of the echo canceler. However, its effect on the common-channel interoffice signaling (CCIS) continuity check between four-wire switching offices has caused some concern.[5]

Common-channel interoffice signaling is a system for exchanging information between processor-equipped switching systems over a network of signaling links. All signaling data for call setup and take-down, as well as network management signals, are exchanged by these systems over the signaling links instead of being sent over the voice path. Thus, a continuity check for voice path assurance (VPA) is conducted whenever a CCIS trunk is selected to switch a call forward. This check not only ensures a satisfactory transmission path but also precludes billing for an otherwise undetectable faulty connection. Between four-wire switching offices the VPA check is performed on a loop basis by connecting a single-frequency transceiver at the originating office and looping the transmission pairs at the distant office. The check is considered successful when the VPA tone received at the originating office is within acceptable transmission and timing limits.

Figure 2 shows the VPA check for the satellite circuit. At the distant office the same notation $x(t)$ is used for the received and the transmitted signals because of the looping configuration. The echo canceler at the distant office will detect double talk during the VPA check. If the echo canceler is not disabled, it may partially cancel the check tone according to the last estimate of the echo path impulse response because of its continued canceling property during double talk. Therefore, the introduction of the echo canceler into the long-distance telephone network without proper disabling control will have some

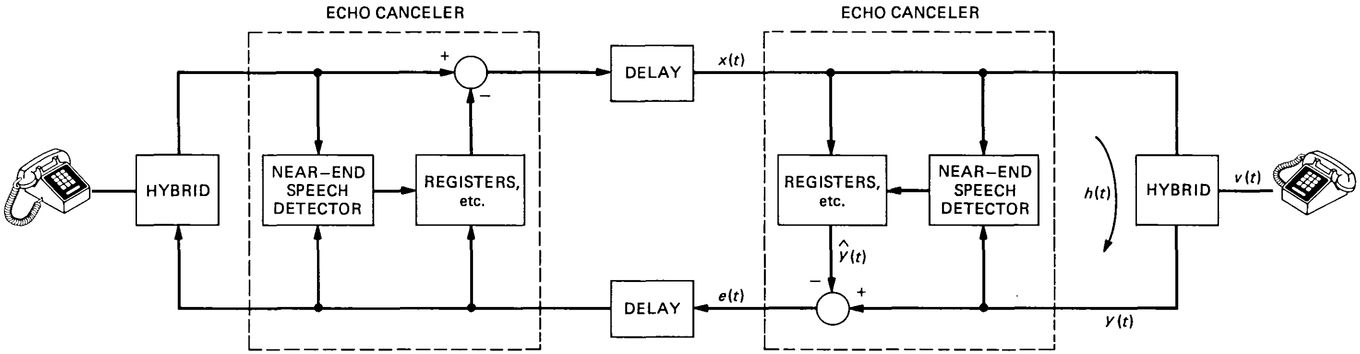


Fig. 1—Use of echo canceler to control echo.

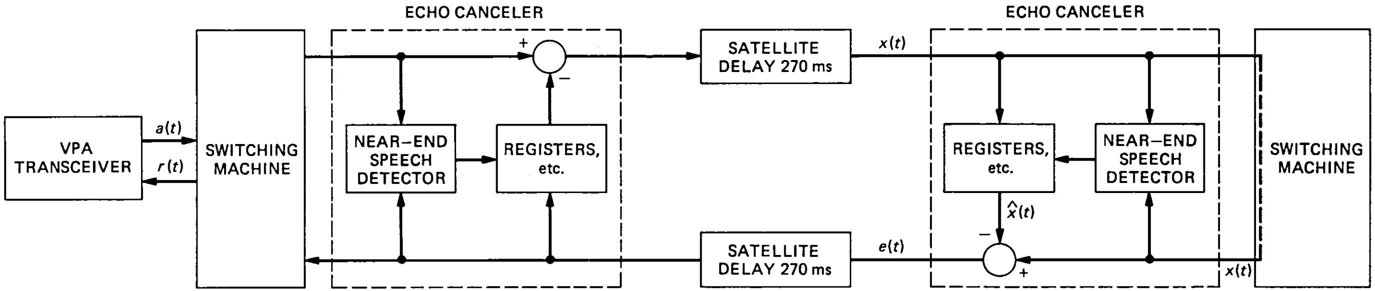


Fig. 2—Echo canceler and CCIS continuity check.

influence on the CCIS continuity check. Voice path assurance failures will lower the performance indices of the affected offices and generate additional maintenance activities. Unfortunately, since call processing information is not readily available, no satisfactory solution has been found such that the echo canceler can be disabled at the desired moment. The seriousness of the echo canceler effect on the VPA check is studied below by evaluating the probabilities of various possible events caused by VPA failures due to the use of the echo canceler. Section II describes the circuits and facilities in the VPA path and how they are modeled mathematically in the derivations given in the Appendices. Section III presents analytic and experimental results and discusses the impact of echo canceler on the CCIS continuity check.

II. MODELING OF VPA CHECK

In this section, the discussion and the derivation will be restricted to No. 4 ESS offices and satellite circuits because the results can be similarly derived for other facilities.

The VPA check for CCIS-equipped No. 4 ESS offices is a two-step test in a specified time interval. The first step is to detect at the proper receive level the transmitted VPA tone which is looped back at the distant office. The second step is to stop transmitting the VPA tone and detect its disappearance at the receiver. At present the specified time interval for completing the test is set anywhere between 2 to 3 s, depending on the specific office. At the beginning of a VPA check during call setup, the originating processor of the switching machine starts the timer and attaches a 2010-Hz transceiver to the selected satellite trunk concurrent with sending an address message through terrestrial CCIS links identifying the trunk to be looped back. The distant office, upon receipt of the message, connects the receive side of the trunk to the transmit side through a zero-loss loop. The originating office checks the level of the returning tone to verify that transmission loss is within acceptable limits. If the returning tone has acceptable level, the originating office stops sending the VPA tone and verifies that the receiver measures below a release level. A VPA failure is generated if for any reason the above two steps are not completed before the timer times out.

The specification of the transceiver levels is given in Table I. Frequent and periodic tests are performed in switching offices to make sure that the transceiver levels are within specification. Table I also gives the test requirements which are stricter than the specification.[6] Since the test only gives pass or fail indication, it is assumed that the various levels tested are uniformly distributed within the test limits. Thus, relative to each other, the transmit and the detect levels of the VPA tone are assumed to be uniformly distributed in (−1, 1) dBm and

Table I—Common-channel interoffice signaling continuity check levels[1]

Test	Specification	Test Requirement
Transmit	$-12 + 1$ dBm0	$-12 \pm$ dBm0
Accept	$-18 + n \leq N$ dBm	$-18.6 + n \leq N$ dBm
Fail	$N \leq -22 + n$ dBm	$N \leq 21.4 + n$ dBm
Release	$N \leq -27 + n$ dBm	—

[1] Where N is the absolute power level of the VPA tone and n is the relative power at the transceiver with respect to the zero transmission level point.

(−9.4, −6.6) dBm, respectively. The release level in Table I is not modeled since a properly working echo canceler should not affect it.

In addition to the transceiver level variations, the VPA tone level is affected by the two-way satellite trunk loss, the check loop, and the echo canceler at the distant office. The satellite trunk is a class of intertoll trunks with 0 dB inserted connected loss which is specified to be normally distributed with a standard deviation of around 0.7 to 0.8 dB. The check loop is specified to have a loss of 0 ± 0.1 dB. This variation is small and will be ignored in the following study. The effect of the echo canceler requires detailed consideration.

In discrete-time notation, the near-end speech detector inhibits updating the echo canceler registers at time k if

$$y(k-l) > \tfrac{1}{2} \max_{0 \leq n \leq 127} x(k-l-n) \qquad \text{for} \quad 0 \leq l \leq 255, \tag{1}$$

where the notations are given in Fig. 1.[3] At an 8-kHz sampling rate, $0 \leq n \leq 127$ implies using a 16-ms tapped delay line adaptive digital filter to model the echo path, while $0 \leq l \leq 255$ means that once near-end speech is detected, the detector continues declaring its presence for the hangover time of 32 ms. The factor of one half is based on the assumption that there will be at least a 6-dB loss through the hybrid. During the continuity check, the VPA tone is looped back so the near-end speech detector finds that eq. (1) is satisfied and thus inhibits updating the registers. The echo canceler will partially cancel the VPA tone according to the register settings determined by the echo path return loss (EPRL) of the last call. Thus, the degree of cancellation is a random variable which is a function of the EPRL. The EPRL distribution of the whole telephone network is not known. The only available EPRL data, as shown in Fig. 3, were measured at the Pittsburgh Regional Center in the second half of 1976 during the field evaluation of domestic satellite.[7] These measurements are not inconsistent with previously reported echo path and intertoll trunk loss.[8] Although the loss distribution was derived from the estimated echo path frequency response as averages over 500 to 2500 Hz, it will be assumed to be the loss distribution for the VPA tone. The density function shown in

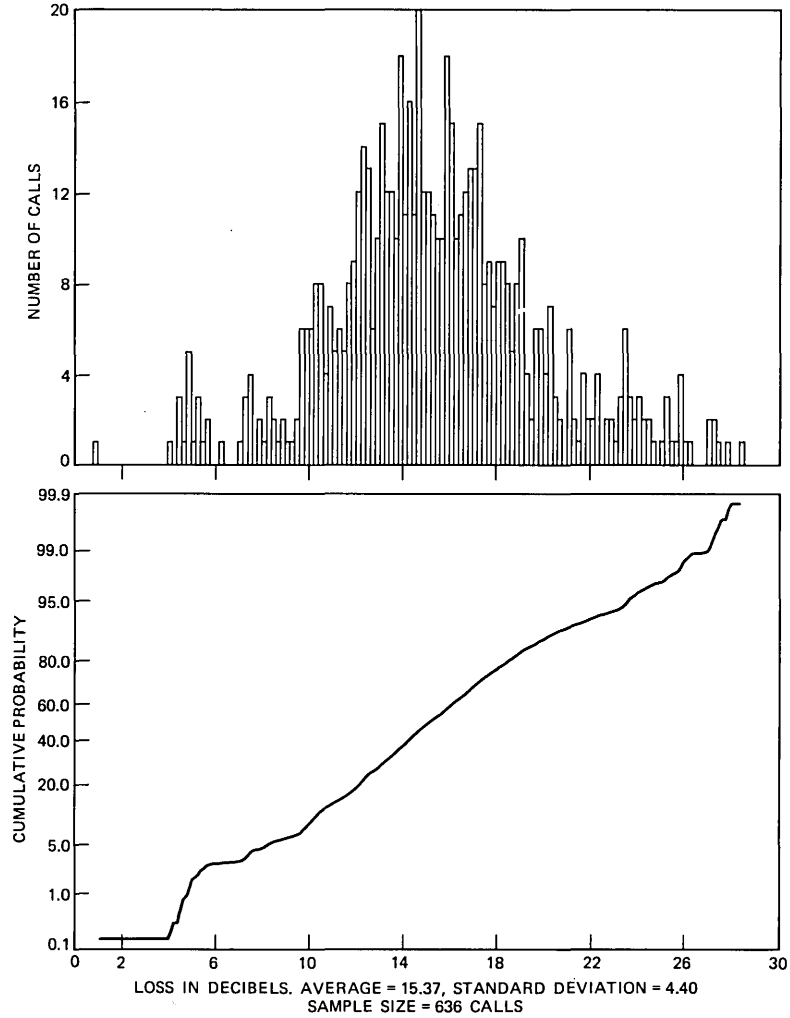


Fig. 3—Distribution of near-end echo path loss, 500 to 2500 Hz flat-weighted.

Fig. 3 is consistent with a normal distribution, except the small secondary peak at about 5 dB. The following derivation turns out to be independent of the loss distribution below 6 dB.

The above modeling of the VPA path can be used to predict VPA failures which may generate various kinds of maintenance activities affecting trunks and carrier groups. There are 12 trunks in each carrier group. If a trunk in a carrier group fails a VPA check, the call is reattempted on another trunk in a different carrier group, while the first carrier group is temporarily taken out-of-service. Several seconds later, a second trunk in the locked-out carrier group is randomly selected and checked for continuity. If the second trunk also fails, a software carrier group alarm (SCGA) is generated for maintenance actions. If the second trunk passes the check, the temporary lockout on the first carrier group is released and the initially failed trunk is rechecked. If it fails again, a single trunk lockout (STL) is generated for maintenance. The probability of getting a VPA failure is derived in Appendix A. The probability of generating an STL is given in Appendix B. The probability of generating SCGA cannot be analyzed since the two failed trunks which share identical facilities from the switching offices to the satellite earth stations may be dependent. Thus, only the upper and the lower bounds of the probability of SCGA are obtained in Appendix C.

III. DISCUSSION OF RESULTS

Figures 4 to 7 show the results obtained in the appendices. During the VPA check, the estimated echo $\hat{x}(t)$ in Fig. 2 is also a single-frequency tone at 2010 Hz. Its amplitude and phase depend on the EPRL of the last call. For the worst-case, 6-dB EPRL, its effect on the looped back tone $x(t)$ varies from 3.5 dB (if the tones add constructively) to −6 dB (if the tones add destructively). With the EPRL distribution given in Fig. 3, the effect of the echo canceler on any VPA tone level follows the density function shown in Fig. 4. This density has a mean value slightly larger than zero; that is, the expected effect of the echo canceler is to increase the VPA tone level slightly. This is

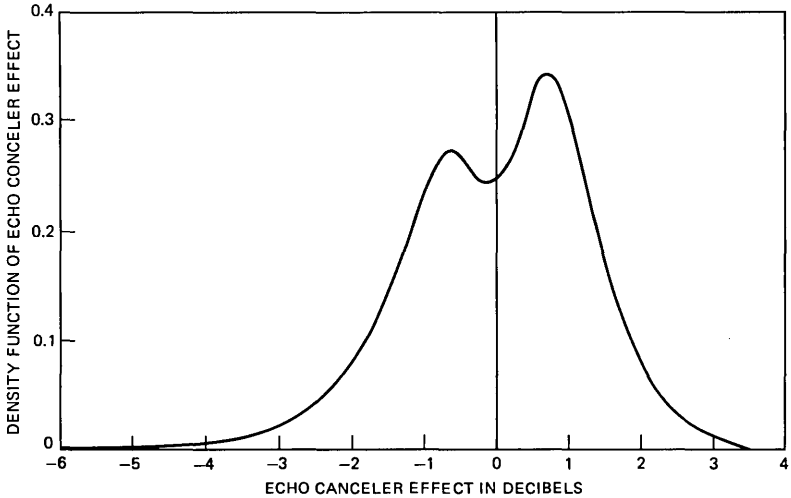


Fig. 4—Probability density function of the effect of the echo canceler on the VPA tone level.

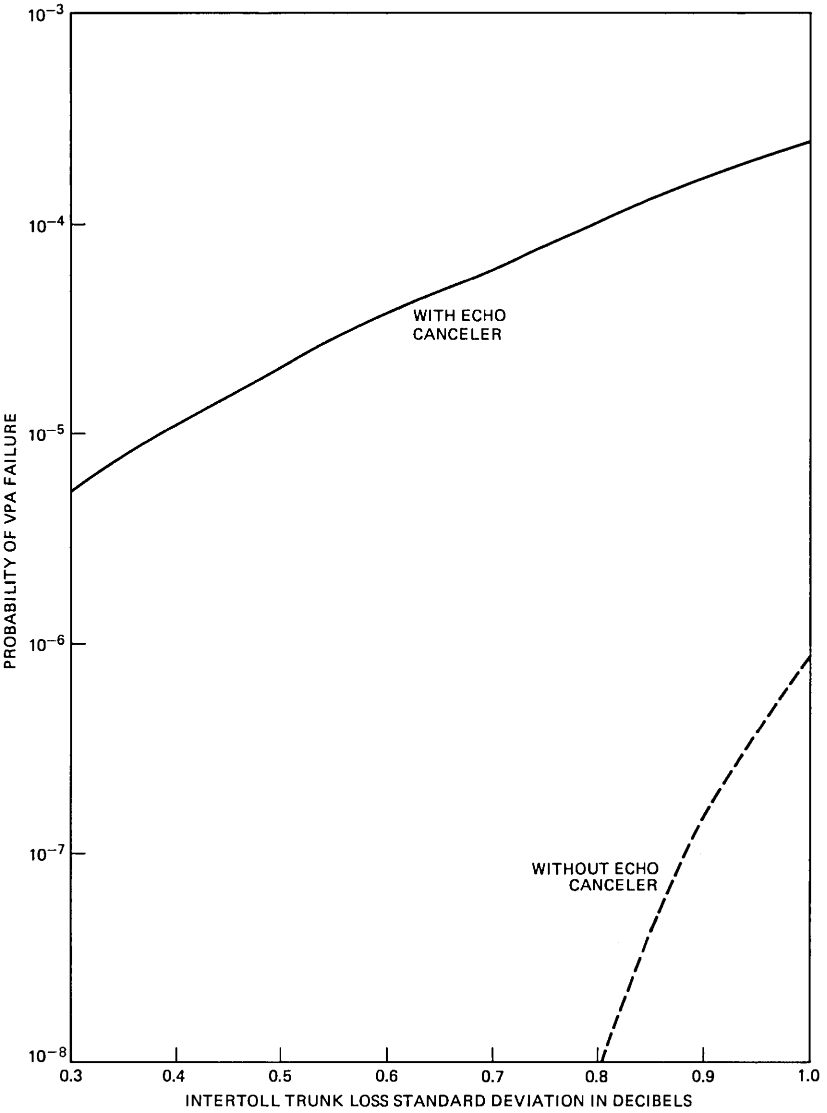


Fig. 5—Probability of VPA failure.

to be expected since the resultant VPA tone contains the powers of two sinusoids $x(t)$ and $\hat{x}(t)$ with random relative phases as opposed to the single sinusoid $x(t)$ without the echo canceler. However, the reducing effect on the VPA tone level has a greater impact because it can be as large as −6 dB. This is evident in Fig. 5, which shows that using echo canceler appreciably increases the probability of VPA failure. However,

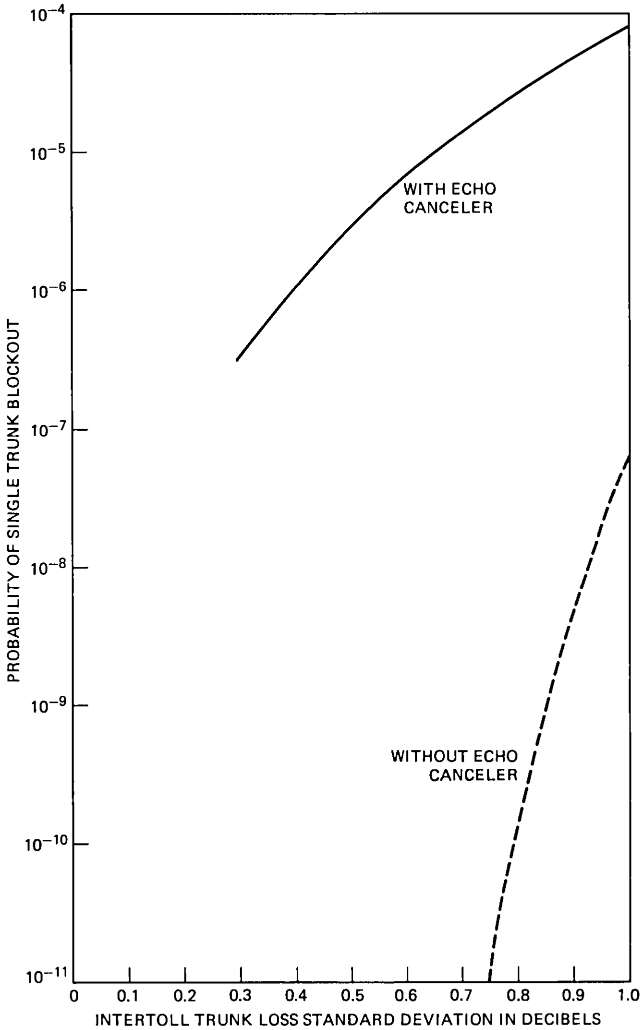


Fig. 6—Probability of single-trunk blockout.

the calculations assume that all circuits are working properly and without outside interferences. Voice path assurance failures can be generated by carrier glitch, fading, and many other causes. Limited sampling performed during the spring of 1980 in several No. 4 ESS offices not equipped with an echo canceler yielded VPA failure rates ranging from 0.5 × 10^{-5} to 2.7 × 10^{-4}, with a sample mean of 5.6 × 10^{-5}.[9] These are of the same order of magnitude as the calculated probability of VPA failure with echo canceler, assuming the satellite trunk standard deviation is between 0.7 to 0.8 dB. Therefore, if the measured VPA failure rate for a group of trunks is 2 × 10^{-5}, it will be probably be around 4 × 10^{-5} after echo cancelers are installed on these trunks. The calculated probabilities of STB and SCGA as given in Figs. 6 and 7, respectively, also do not appear to be excessive, although there is no field statistics for comparison.

An experiment was performed on the full-hop satellite circuits between Atlanta, Georgia, and Cedar Knolls, New Jersey to see the effect of not disabling the echo canceler prior to the VPA check. The data indicated that there was no VPA failure for well over 100,000 calls.

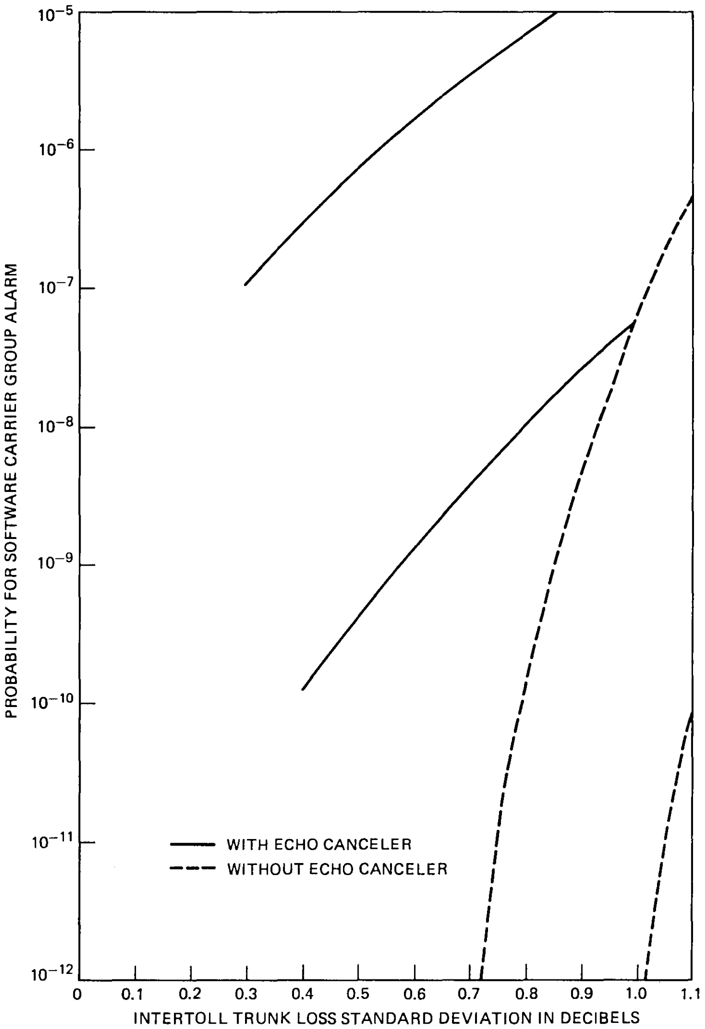


Fig. 7—Probability bounds for software carrier group alarm.

This result is better than the calculated probabilities and it can be easily explained. The calculated numbers are based on the specifications of the trunks and circuits. The standard deviation of the satellite trunks appears to be smaller than that of the intertoll trunks. Furthermore, limited measurements of the actual transceiver levels indicated that the transmit and the detect levels can be modeled as uniformly distributed in (−0.25, 0.25) dBm and (−8.75, −7.75) dBm, respectively.[10] All these factors help the performance of the VPA check.

In conclusion, the introduction of echo canceler to the satellite circuits may generate additional VPA failures. However, in present long distance networks, they are not significant compared to existing VPA failure statistics of the field. Thus, it does not appear necessary to disable the echo canceler prior to each CCIS continuity check.

APPENDIX A

The probability of VPA failure is derived in this appendix. Let l be the EPRL shown in Fig. 3; that is, l is normal with mean $\mu = 15.37$ and standard deviation $\sigma = 4.4$, except for $l < 6$ dB. By definition, for a single frequency tone,

$$\begin{aligned} l &\approx 10 \log \frac{\text{power of } x(t)}{\text{power of } y(t)} \\ &= 20 \log \frac{\text{amplitude of } x(t)}{\text{amplitude of } y(t)} \\ &\approx 20 \log \frac{\text{amplitude of } x(t)}{\text{amplitude of } \hat{y}(t)}, \end{aligned}$$

where the last step follows for $l > 6$ dB. For the loopback configuration in Fig. 2 during the VPA check, then

$$l \approx 20 \log \frac{1}{b}, \qquad l > 6 \text{ dB}, \tag{1}$$

where

$$b = \frac{\text{amplitude of } \hat{x}(t)}{\text{amplitude of } x(t)}.$$

The derivation below assumes that eq. (1) is an exact equality. Since $l > 6$ dB, b is distributed in (0, ½). Let l_i, $i = 1, 2, \cdots$, denote the EPRL of the ith previous call using a specific echo canceler. The l_i's are identically distributed as l, and for most practical situations, independent random variables. Clearly b is a function of the l_i's. For instance, if $l_1 > 6$ dB, b is determined solely by l_1. If $l_1 < 6$ dB, the near-end speech detector would declare double talk and no register update would take place during the last call. Then b would be independent of

l_1 and become a function of only l_i, $i = 2, 3, \cdots$. Thus, the distribution function of b is given by

$$\begin{aligned} P\{b < c\} &= P\{10^{-l_1/20} < c\} + P\{l_1 < 6\}P\{10^{-l_2/20} < c\} \\ &\qquad + P\{l_1 < 6\}P\{l_2 < 6\}P\{10^{-l_3/20} < c\} + \cdots \\ &= (1 + (1 - g) + (1 - g)^2 + \cdots)P\{10^{-l/20} < c\} \\ &= \frac{1}{g} P\left\{ l > 20 \log \frac{1}{c} \right\} \\ &= \frac{1}{g} \int_{20 \log 1/c}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-(l-\mu)^2/2\sigma^2} dl, \end{aligned} \tag{2}$$

where

$$\begin{aligned} g &= P\{l > 6\} \\ &= \int_6^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp[-(l - \mu)^2/2\sigma^2]\, dl. \end{aligned}$$

Differentiating eq. (2) gives the density function of b as

$$f_b(b) = \frac{20}{\sqrt{2\pi}\sigma g \ln 10} \frac{1}{b} \exp[-(20 \log b + \mu)^2/2\sigma^2], \qquad 0 < b < \frac{1}{2}. \tag{3}$$

The transmitted VPA tone $a(t)$ can be written as

$$a(t) = \sqrt{2}A \sin \omega t.$$

The transmit level in dBm

$$x_1 = 10 \log \frac{A^2}{10^{-3}}$$

is uniformly distributed in $(-1, 1)$, as described in Section II. Since the satellite trunk loss in dB is normally distributed, it can be considered as gain for convenience. Assume the satellite trunk gains in the transmit and the receive directions are denoted by g_1 and g_2, respectively. Let

$$x_2 = 10 \log g_1^2$$

and

$$x_3 = 10 \log g_2^2.$$

The terms x_2 and x_3 are independent normal random variables with zero mean and standard deviation ρ. Thus, using the notation of Fig. 2,

$$\begin{aligned} x(t) &= \sqrt{2}Ag_1 \sin \omega t, \\ e(t) &= \sqrt{2}Ag_1 \sin \omega t - \sqrt{2}Ag_1 b \sin(\omega t + \phi) \\ &= \sqrt{2}Ag_1 \sqrt{1 + b^2 - 2b \cos \phi} \cos \left[\omega t + \tan^{-1} \left(\frac{1 - b \cos \phi}{b \sin \phi} \right) \right], \end{aligned}$$

where ϕ, the relative phase between the loopback signal $x(t)$ and the echo estimate $\hat{x}(t)$, is assumed to be uniformly distributed in $(0, 2\pi)$. The signal received by the transceiver

$$\gamma(t) = g_2 e(t)$$

has a power level in dBm of

$$\begin{aligned} \gamma &= 10 \log \frac{A^2 g_1^2 g_2^2}{10^{-3}} (1 + b^2 - \cos \phi) \\ &= x_1 + x_2 + x_3 + w, \end{aligned} \tag{4}$$

where

$$w = 10 \log(1 + b^2 - 2b \cos \phi) \tag{5}$$

is the contribution of the echo canceler to the receive level of the VPA tone. Its effect can be evaluated by deriving the distribution of w. Since the density functions of b and ϕ are known, the density function of w can be shown to be

$$f_w(w) = \begin{array}{ll} f_{w1}(w), & 10 \log 0.25 < w < 10 \log 0.75, \\ f_{w2}(w), & 10 \log 0.75 < w < 0, \\ f_{w3}(w), & 0 < w < 10 \log 2.25, \end{array} \tag{6}$$

where

$$\begin{aligned} f_{w1}(w) &= k_1 \int_{1.25-q_3}^{1} \frac{q_3 q_5}{q_4 q_2}\, dv, \\ f_{w2}(w) &= k_1 \int_{\sqrt{1-q_3}}^{1} \frac{q_3 q_5}{q_4 q_2}\, dv + k_1 \int_{\sqrt{1-q_3}}^{1.25-q_3} \frac{q_3 q_6}{q_4 q_1}\, dv, \\ f_{w3}(w) &= k_1 \int_{-1}^{1.25-q_3} \frac{q_3 q_6}{q_4 q_1}\, dv, \\ k_1 &= \frac{1}{\sqrt{2\pi}\pi\sigma g}, \\ q_1 &= v + \sqrt{v^2 - 1 + q_3}, \\ q_2 &= v - \sqrt{v^2 - 1 + q_3}, \\ q_3 &= 10^{w/10}, \\ q_4 &= \sqrt{(v^2 - 1 + q_3)(1 - v^2)}, \\ q_5 &= \exp[-(20 \log q_2 + \mu)^2 / 2\sigma^2], \end{aligned} \tag{7}$$

and

$$q_6 = \exp[-(20 \log q_1 + \mu)^2/2\sigma^2].$$

The density function $f_w(w)$ is plotted in Fig. 4. Since the mean value of $1 + b^2 - 2b \cos \phi$ is slightly larger than one, the expected value of w is slightly larger than zero.

The receive level γ in eq. (4) is the sum of random variables with known density functions; therefore, its distribution is also known. The transceiver detector level u is assumed to be uniformly distributed in $(-9.4, -6.6)$ dBm. Thus, the probability of VPA failure with the echo canceler can be derived as

$$P_{ve} = P\{\gamma < u\}$$

$$= \frac{k_1}{11.2}\int_{-\infty}^{0}\Bigg\{\int_{10\log 0.25}^{10\log 0.75}\int_{z-w-9.4}^{z-w-6.6} q_7\, dx \int_{1.25-q_3}^{1} \frac{q_3 q_5}{q_4 q_2}\, dv\, dw$$

$$+ \int_{10\log 0.75}^{0}\int_{z-w-9.4}^{z-w-6.6} q_7\, dx \left(\int_{\sqrt{1-q_3}}^{1} \frac{q_3 q_5}{q_4 q_2}\, dv + \int_{\sqrt{1-q_3}}^{1.25-q_3} \frac{q_3 q_6}{q_4 q_1}\, dv\right) dw$$

$$+ \int_{0}^{10\log 2.25}\int_{z-w-9.4}^{z-w-6.6} q_7\, dx \int_{1}^{1.25-q_3} \frac{q_3 q_6}{q_4 q_1}\, dv\, dw\Bigg\}\, dz. \qquad (8)$$

The probability of VPA failure without the echo canceler can be calculated as

$$P_v = P\{x_1 + x_2 + x_3 < u\} = \frac{1}{11.2}\int_{-\infty}^{0}\int_{z-9.4}^{z-6.6} q_7\, dx\, dz, \qquad (9)$$

where

$$q_7 = \operatorname{erf}\left(\frac{x+1}{2\rho}\right) - \operatorname{erf}\left(\frac{x-1}{2\rho}\right)$$

and

$$\operatorname{erf}(x) = \frac{2}{\sqrt{2\pi}}\int_{0}^{\sqrt{2}x} e^{-y^2/2}\, dy$$

is the error function. Equations (8) and (9) are used to plot Fig. 5.

APPENDIX B

The probability of STL is derived in this Appendix. An STL is generated on a trunk if the trunk fails a VPA check, if another trunk in the same carrier group passes a subsequent VPA check, and if the first trunk fails a VPA recheck. For simplicity, the second trunk is assumed to be independent of the first trunk. The probability that a trunk fails two consecutive VPA checks will be derived first. It is assumed that the small probability of selecting the same transceiver for the two VPA checks is negligible.

A VPA check on a trunk fails if $x_1 + x_2 + x_3 + w < u$. The values of x_2, x_3, and w remains unchanged for a VPA recheck on the same trunk. Thus, the probability of interest is $P\{x_2 + x_3 + w < u - x_1\}$. Let $z = u - x_1$, $s = x_2 + x_3$ and $y = x_2 + x_3 + w$. It can be shown that

$$f_z(z) = \begin{cases} \dfrac{1}{5.6}(10.4 + z), & -8.4 \geq z \geq -10.4, \\ \dfrac{2}{5.6}, & -7.6 \geq z \geq -8.4, \\ \dfrac{1}{5.6}(-z - 5.6), & -5.6 \geq z \geq -7.6, \end{cases}$$

$$f_s(s) = \frac{1}{2\sqrt{\pi}\rho} \exp(-s^2/4\rho^2), \tag{10}$$

and

$$\begin{aligned} f_y(y) = {} & k_2 \int_{10\log 0.25}^{10\log 0.75} \int_{1.25-q_3}^{1} \frac{q_3 q_8 q_5}{q_4 q_2}\, dv\, dw \\ & + k_2 \int_{10\log 0.75}^{0} \left\{ \int_{\sqrt{1-q_3}}^{1} \frac{q_3 q_8 q_5}{q_4 q_2}\, dv + \int_{\sqrt{1-q_3}}^{1.25-q_3} \frac{q_3 q_8 q_6}{q_4 q_1}\, dv \right\} \\ & + k_2 \int_{0}^{10\log 2.25} \int_{-1}^{1.25-q_3} \frac{q_3 q_8 q_6}{q_4 q_1}\, dv\, dw, \end{aligned}$$

where

$$k_2 = \frac{1}{2\sqrt{2}\pi^2 \sigma \rho g}$$

and

$$q_8 = \exp[-(y - w)^2/4\rho^2].$$

For a given value of y, a VPA failure is generated with the probability

$$P\{y < z\} = \int_{y}^{\infty} f_z(z)\, dz$$

$$= \begin{cases} 1, & y \leq -10.4, \\ -\dfrac{1}{5.6}\left(48.48 + 10.4y + \dfrac{y^2}{2}\right), & -8.4 \geq y \geq -10.4, \\ -\dfrac{1}{5.6}(13.2 + 2y), & -7.6 \geq y \geq -8.4, \\ \dfrac{1}{5.6}\left(15.68 + 5.6y + \dfrac{y^2}{2}\right), & -5.6 \geq y \geq -7.6, \\ 0, & y \geq -5.6. \end{cases}$$

Therefore, the probability of two consecutive VPA failures on a trunk with echo canceler is

$$P_{2e} = \int_{-\infty}^{\infty} P\{y < z\}P\{y < z\} f_y(y)\, dy$$

$$= \int_{-\infty}^{-10.4} f_y(y)\, dy + \frac{1}{5.6^2}\int_{-10.4}^{-8.4}\left(48.48 + 10.4y + \frac{y^2}{2}\right)^2$$

$$\times f_y(y)\, dy + \frac{1}{5.6^2}\int_{-8.4}^{-7.6} (13.2 + 2y)^2 f_y(y)\, dy$$

$$+ \frac{1}{5.6^2}\int_{-7.6}^{-5.6}\left(15.68 + 5.6y + \frac{y^2}{2}\right)^2 f_y(y)\, dy. \tag{11}$$

The probability of STL with echo canceler is then

$$P_{te} = P_{2e} \cdot (1 - P_{ve}). \tag{12}$$

Without echo canceler the probability of two consecutive VPA failures on a trunk, P_2, is obtained by replacing $f_y(y)$ in eq. (11) by $f_s(y)$ in (10). The probability of STL without echo canceler is then

$$P_t = P_2 \cdot (1 - P_v). \tag{13}$$

Equations (12) and (13) are used to plot Fig. 6.

APPENDIX C

The probability of SCGA is studied in this appendix. An SCGA is generated if two trunks in the same carrier group fail successive VPA checks. If the two trunks are independent, the probability of SCGA is simply the square of the probability of VPA failure. This is plotted as lower bounds in Fig. 7. Since the two trunks share identical carrier facilities from the switching offices to the earth stations, they may be dependent in some unknown way. A pessimistic estimate of the probability of SCGA is to assume that x_2 and x_3 remain the same for the two trunks selected. Thus, the probability of itnerest is $P\{x_1 + w - u < -(x_2 + x_3)\}$.

Let $t = x_1 + w$ and $h = t - u$. Then,

$$f_t(t) = \begin{cases} f_{t1}(t), & 10 \log 2.25 - 1 < t < 10 \log 2.25 + 1, \\ f_{t2}(t), & 1 < t < 10 \log 2.25 - 1, \\ f_{t3}(t), & 10 \log 0.75 + 1 < t < 1, \\ f_{t4}(t), & -1 < t < 10 \log 0.75 + 1, \\ f_{t5}(t), & 10 \log 0.75 - 1 < t < -1, \\ f_{t6}(t), & 10 \log 0.25 + 1 < t < 10 \log 0.75 - 1, \\ f_{t7}(t), & 10 \log 0.25 - 1 < t < 10 \log 0.25 + 1, \end{cases}$$

where

$$f_{t1}(t) = \frac{1}{2} \int_{t-1}^{10 \log 0.25} f_{w3}(w)\, dw,$$

$$f_{t2}(t) = \frac{1}{2} \int_{t-1}^{t+1} f_{w3}(w)\, dw,$$

$$f_{t3}(t) = \frac{1}{2} \int_{0}^{t+1} f_{w3}(w)\, dw + \frac{1}{2} \int_{t-1}^{0} f_{w2}(w)\, dw,$$

$$f_{t4}(t) = \frac{1}{2} \int_{0}^{t+1} f_{w3}(w)\, dw + \frac{1}{2} \int_{10 \log 0.75}^{0} f_{w2}(w)\, dw + \frac{1}{2} \int_{t-1}^{10 \log 0.75} f_{w1}(w)\, dw,$$

$$f_{t5}(t) = \int_{10 \log 0.75}^{t+1} f_{w2}(w)\, dw + \frac{1}{2} \int_{t-1}^{10 \log 0.75} f_{w1}(w)\, dw,$$

$$f_{t6}(t) = \frac{1}{2} \int_{t-1}^{t+1} f_{w1}(w)\, dw,$$

and

$$f_{t7}(t) = \frac{1}{2} \int_{10 \log 0.25}^{t+1} f_{w1}(w)\, dw.$$

The density function of h is given by

$$f_h(h) = \begin{cases} f_{h1}(h), & 10 \log 2.25 + 8.4 < h < 10 \log 2.25 + 10.4, \\ f_{h2}(h), & 10 \log 2.25 + 7.6 < h < 10 \log 2.25 + 8.4, \\ f_{h3}(h), & 10.4 < h < 10 \log 2.25 + 7.6, \\ f_{h4}(h), & 10 \log 0.75 + 10.4 < h < 10.4, \\ f_{h5}(h), & 10 \log 0.75 + 5.6 < h < 10 \log 0.75 + 10.4, \\ f_{h6}(h), & 8.4 < h < 10 \log 2.25 + 5.6, \\ f_{h7}(h), & 7.6 < h < 8.4, \\ f_{h8}(h), & 10 \log 0.75 + 8.4 < h < 7.6, \\ f_{h9}(h), & 10 \log 0.75 + 7.6 < h < 10 \log 0.75 + 8.4, \\ f_{h10}(h), & 5.6 < h < 10 \log 0.75 + 7.6, \\ f_{h11}(h), & 10 \log 0.25 + 10.4 < h < 5.6, \\ f_{h12}(h), & 10 \log 0.75 + 5.6 < h < 10 \log 0.25 + 10.4, \\ f_{h13}(h), & 10 \log 0.25 + 8.4 < h < 10 \log 0.75 + 5.6, \\ f_{h14}(h), & 10 \log 0.25 + 7.6 < h < 10 \log 0.25 + 8.4, \\ f_{h15}(h), & 10 \log 0.25 + 5.6 < h < 10 \log 0.25 + 7.6, \end{cases}$$

where

$$f_{h1}(h) = \frac{1}{2.8} \int_{h-9.4}^{10\log 2.25+1} f_{t1}(t)\, dt,$$

$$f_{h2}(h) = \frac{1}{2.8} \int_{10\log 2.25-1}^{10\log 2.25+1} f_{t1}(t)\, dt + \frac{1}{2.8} \int_{h-9.4}^{10\log 2.25-1} f_{t2}(t)\, dt,$$

$$f_{h3}(h) = \frac{1}{2.8} \int_{10\log 2.25-1}^{h-6.6} f_{t1}(t)\, dt + \frac{1}{2.8} \int_{h-9.4}^{10\log 2.25-1} f_{t2}(t)\, dt,$$

$$f_{h4}(h) = \frac{1}{2.8} \int_{10\log 2.25-1}^{h-6.6} f_{t1}(t)\, dt + \frac{1}{2.8} \int_{1}^{10\log 2.25-1} f_{t2}(t)\, dt$$
$$+ \frac{1}{2.8} \int_{h-9.4}^{1} f_{t3}(t)\, dt,$$

$$f_{h5}(h) = \frac{1}{2.8} \int_{10\log 2.25-1}^{h-6.6} f_{t1}(t)\, dt + \frac{1}{2.8} \int_{1}^{10\log 2.25-1} f_{t2}(t)\, dt$$
$$+ \frac{1}{2.8} \int_{10\log 0.75+1}^{1} f_{t3}(t)\, dt + \frac{1}{2.8} \int_{h-9.4}^{10\log 0.75+1} f_{t4}(t)\, dt,$$

$$f_{h6}(h) = \frac{1}{2.8}\int_{1}^{h-6.6} f_{t2}(t)\,dt + \frac{1}{2.8}\int_{10\log 0.75+1}^{1} f_{t3}(t)\,dt$$
$$+ \frac{1}{2.8}\int_{h-9.4}^{10\log 0.75+1} f_{t4}(t)\,dt,$$

$$f_{h7}(h) = \frac{1}{2.8}\int_{1}^{h-6.6} f_{t2}(t)\,dt + \frac{1}{2.8}\int_{10\log 0.75+1}^{1} f_{t3}(t)\,dt$$
$$+ \frac{1}{2.8}\int_{-1}^{10\log 0.75+1} f_{t4}(t)\,dt + \frac{1}{2.8}\int_{h-9.4}^{-1} f_{t5}(t)\,dt,$$

$$f_{h8}(h) = \frac{1}{2.8}\int_{10\log 0.75+1}^{h-6.6} f_{t3}(t)\,dt + \frac{1}{2.8}\int_{-1}^{10\log 0.75+1} f_{t4}(t)\,dt$$
$$+ \frac{1}{2.8}\int_{h-9.4}^{-1} f_{t5}(t)\,dt,$$

$$f_{h9}(h) = \frac{1}{2.8}\int_{10\log 0.75+1}^{h-6.6} f_{t3}(t)\,dt + \frac{1}{2.8}\int_{10\log 0.75-1}^{-1} f_{t5}(t)\,dt$$
$$+ \frac{1}{2.8}\int_{-1}^{10\log 0.75+1} f_{t4}(t)\,dt + \frac{1}{2.8}\int_{h-9.4}^{10\log 0.75-1} f_{t6}(t)\,dt,$$

$$f_{h10}(h) = \frac{1}{2.8}\int_{-1}^{h-6.6} f_{t4}(t)\,dt + \frac{1}{2.8}\int_{10\log 0.75-1}^{-1} f_{t5}(t)\,dt$$
$$+ \frac{1}{2.8}\int_{h-9.4}^{10\log 0.75-1} f_{t6}(t)\,dt,$$

$$f_{h11}(h) = = \frac{1}{2.8}\int_{10\log 0.75-1}^{h-6.6} f_{t5}(t)\,dt + \frac{1}{2.8}\int_{h-9.4}^{10\log 0.75-1} f_{t6}(t)\,dt,$$

$$f_{h12}(h) = \frac{1}{2.8}\int_{10\log 0.75-1}^{h-6.6} f_{t5}(t)\,dt + \frac{1}{2.8}\int_{10\log 0.25+1}^{10\log 0.75-1} f_{t6}(t)\,dt$$
$$+ \frac{1}{2.8}\int_{h-9.4}^{10\log 0.25+1} f_{t7}(t)\,dt,$$

$$f_{h13}(h) = \frac{1}{2.8}\int_{10\log 0.25+1}^{h-6.6} f_{t6}(t)\,dt + \frac{1}{2.8}\int_{h-9.4}^{10\log 0.25+1} f_{t7}(t)\,dt$$

$$f_{h14}(h) = \frac{1}{2.8}\int_{10\log 0.25+1}^{h-6.6} f_{t6}(t)\,dt + \frac{1}{2.8}\int_{10\log 0.25-1}^{10\log 0.25+1} f_{t7}(t)\,dt,$$

$$f_{h15}(h) = \frac{1}{2.8}\int_{10\log 0.25-1}^{h-6.6} f_{t7}(t)\,dt.$$

For a given value of s, a VPA failure is generated with the probability

$$P\{h < -s\} = \int_{-\infty}^{-s} f_h(h)\,dh$$

$$= \left\{ \begin{array}{l}
0, \qquad \text{for} \quad -s < 10\log 0.25 + 5.6, \\[2ex]
\displaystyle\int_{10\log 0.25+5.6}^{-s} f_{h15}(h)\,dh, \\[2ex]
\quad \text{for} \quad 10\log 0.25 + 5.6 < -s < 10\log 0.25 + 7.6, \\[2ex]
\displaystyle\int_{10\log 0.25+5.6}^{10\log 0.25+7.6} f_{h15}(h)\,dh + \int_{10\log 0.25+7.6}^{-s} f_{h14}(h)\,dh, \\[2ex]
\quad \text{for} \quad 10\log 0.25 + 7.6 < -s < 10\log 0.25 + 8.4, \\[2ex]
\displaystyle\int_{10\log 0.25+5.6}^{10\log 0.25+7.6} f_{h15}(h)\,dh + \int_{10\log 0.25+7.6}^{10\log 0.25+8.4} f_{h14}(h)\,dh \\[2ex]
\quad \displaystyle + \int_{10\log 0.25+8.4}^{-s} f_{h13}(h)\,dh, \\[2ex]
\quad \text{for} \quad 10\log 0.25 + 8.4 < -s < 10\log 0.75 + 5.6, \\[2ex]
\displaystyle\int f_{h15} + \int f_{h14} + \int_{10\log 0.25+8.4}^{10\log 0.75+5.6} f_{h13}(h)\,dh \\[2ex]
\quad \displaystyle + \int_{10\log 0.75+5.6}^{-s} f_{h12}(h)\,dh, \\[2ex]
\quad \text{for} \quad 10\log 0.75 + 5.6 < -s < 10\log 0.25 + 10.4 \\[1ex]
\qquad\qquad\qquad\qquad \vdots \\[1ex]
\displaystyle\int f_{h15} + \int f_{h14} + \cdots + \int_{10\log 2.25+7.6}^{10\log 2.25+8.4} f_{h2}(h)\,dh \\[2ex]
\quad \displaystyle + \int_{10\log 2.25+8.4}^{-s} f_{h1}(h)\,dh, \\[2ex]
\quad \text{for} \quad 10\log 2.25 + 8.4 < -s < 10\log 2.25 + 10.4 \\[2ex]
1, \qquad \text{for} \quad 10\log 2.25 + 10.4 < -s.
\end{array} \right.$$

Let $p = -s$. The pessimistic estimate of the probability of SCGA with echo canceler is given by

$$\begin{aligned}
&\int_{-\infty}^{\infty} P\{h<p\}P\{h<p\}\, f_s(p)\, dp \\
&= \int_{10\log 0.25+5.6}^{10\log 0.25+7.6} \Bigg(\int_{10\log 0.25+5.6}^{p} f_{h15}(h)\, dh\Bigg)^2 f_s(p)\, dp \\
&\quad + \int_{10\log 0.25+7.6}^{10\log 0.25+8.4} \Bigg(\int_{10\log 0.25+5.6}^{10\log 0.25+7.6} f_{h15}(h)\, dh \\
&\quad + \int_{10\log 0.25+7.6}^{p} f_{h14}(h)\, dh\Bigg)^2 f_s(p)\, dp \\
&\quad + \cdots + \int_{10\log 2.25+8.4}^{10\log 2.25+10.4} \Bigg(\int f_{h15} + \int f_{h14} + \cdots \\
&\quad + \int_{10\log 2.25+7.6}^{10\log 2.25+8.4} f_{h2}(h)\, dh + \int_{10\log 2.25+8.4}^{p} f_{h1}(h)\, dh\Bigg)^2 f_s(p)\, dp \\
&\quad + \int_{10\log 2.25+10.4}^{\infty} (5.6)^2 f_s(p)\, dp. \qquad (14)
\end{aligned}$$

Without echo canceler, the pessimistic estimate of the probability of SCGA is simply P_t, the probability of STL given in eq. (13). Equations (13) and (14) are used to plot the upper bounds of Fig. 7.

REFERENCES

1. *Transmission Systems for Communications,* Technical Staff of Bell Telephone Laboratories, Inc., Western Electric Company: Winston-Salem, North Carolina, 1970, 4th ed.
2. M. M. Sondhi and D. A. Berkeley, "Silencing Echoes on the Telephone Network," Proc. IEEE, *68,* No. 8 (August 1980), pp. 948–63.
3. D. L. Duttweiler, "A Twelve-Channel Digital Echo Canceler," IEEE Trans. Commun., *COM-26* (May 1978), pp. 647–53.
4. D. L. Duttweiler and Y. S. Chen, "A Single-Chip VLSI Echo Canceler," B.S.T.J., *59,* No. 2 (February 1980), pp. 149–60.
5. Special Issue on Common Channel Interoffice Signaling, B.S.T.J., *57,* No. 2 (February 1978).
6. H. A. Kerr, private communication.
7. M. B. Brilliant, unpublished.
8. F. P. Duffy et al., "Echo Performance of Toll Telephone Connections in the United States," B.S.T.J., *54,* No. 2 (February 1975), pp. 209–43.
9. E. A. Davis, private communication.
10. R. Metz, private communication.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 60, No. 7, September 1981
Printed in U.S.A.

Profile Characterization of Optical Fibers—A Comparative Study

By H. M. PRESBY

(Manuscript received February 12, 1981)

The refractive index profiles of several multimode optical fibers were measured by four of the current state-of-the-art techniques. These include interference microscopy on slab samples, transverse interference microscopy on whole fiber samples, the focusing method, and the refracted near-field method. The profile of the parent preform of one of the fibers was also measured by both the focusing method and by a ray-tracing approach. Comparisons of the results and the measurement methods are made emphasizing the applicability and use of the various techniques.

I. INTRODUCTION

Precise methods for measuring index profiles in both multimode and single-mode optical fibers and preforms are required if the desired ideal index distributions are to be produced. This paper will provide a comparison of four current state-of-the-art techniques for making these measurements on fibers and two similarly current methods used for preforms. Implications of the results especially related to the applicability and utilization of the various techniques will be discussed.

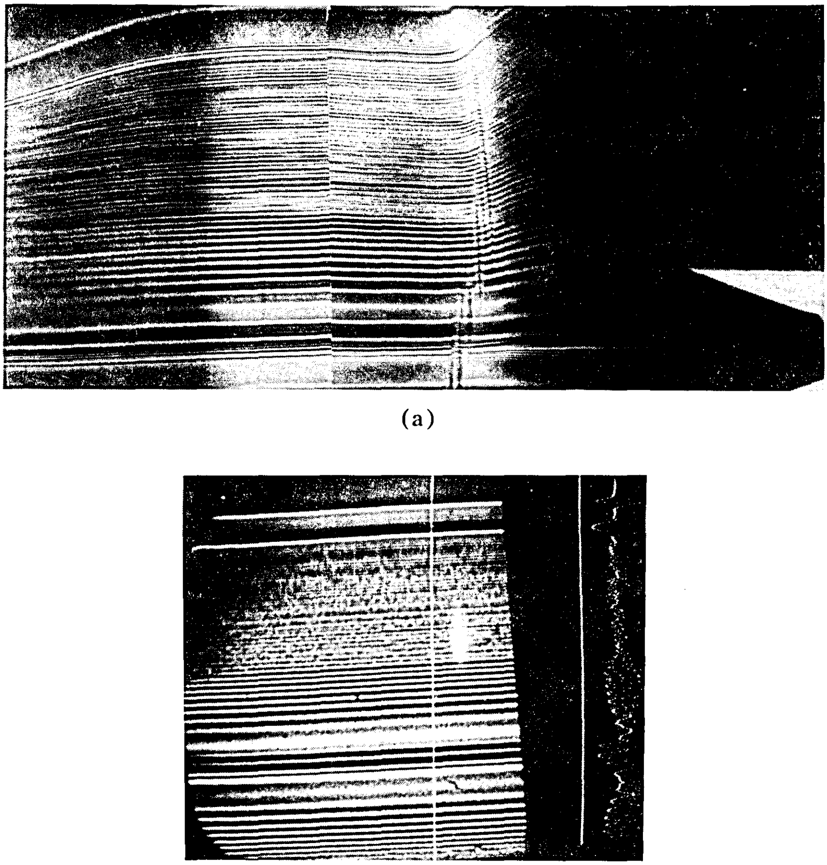
It is important to have some feeling for the structure of the object being profiled. The fibers and the preform studied were produced by the modified chemical vapor deposition process.[1] In this procedure a silica tube is mounted on a glass-working lathe and slowly rotated while reactants and dopants flow through it in an oxygen stream. An oxy-hydrogen burner is slowly traversed along the outside of the tube to provide simultaneous deposition and fusion of a layer of the reacting materials. On the order of 50 layers are deposited by multiple passes of the burner. To fabricate graded-index fibers, the dopant concentration is gradually increased with increasing layer number.

At the conclusion of deposition, the temperature of the burner is raised to collapse the tube into a solid preform. A diagnostigram[2] of the remaining end of the preform after being pulled into a fiber is shown in Fig. 1a. This display is obtained by expanding and collimating the light from a CW laser and allowing it to impinge upon the preform. The diagnostigram provides a simple and rapid nondestructive means of investigating internal layer structure. Details of its operation can be found in the Appendix. A shadow of the blunted end of the preform is seen at the lower right and each of the deposition layers is seen as a horizontal line. The uppermost bright line is the core-cladding boundary and the region immediately below it is a 320-μm-thick barrier layer. The core radius is about 3.5 mm and the length of preform seen in the diagnostigram is 12 cm. The point to note is the rich structure and resolvability of the deposition layers. Another view of this structure can be obtained by immersing the preform in index-matching oil and observing the incoherently illuminated core with a video camera, as is done in the focusing method.[3] This representation, shown in Fig. 1b for a several millimeter length of preform, also emphasizes the individual layer structure, which is in addition displayed by the curve at the side of the display.

These structural variations also appear, appropriately scaled, in the fiber in a one-to-one correspondence[4] and corroborates the fact that the same distribution of refractive index that is introduced into the preform exists in the fiber.[5] Generally, the scale of the variations in the fiber is on the order of less than a wavelength and they are, therefore, not observed by the measurement technique. Notable exceptions occur near the axis where the deposition layers are thickest and in any region where either several layers have the same index or thicker than normal layers are produced, due to fabrication faults.

The profile, built up by varying the dopant concentration in each of the layers, can be measured by a variety of techniques. How accurately the measurements should be made and which technique is best for making them are always difficult questions to answer since they involve trade-offs of many factors. To be included are time, cost, and effort considerations, all of which one would like to minimize, and sensitivity, accuracy, and resolution which one would like to maximize.

A handle on the question of accuracy is provided by the following considerations. The theoretical bandwidth that can be realized with an optimum profile is about 8000 MHz × km for a fiber with a maximum index difference value of 0.02.[6] To achieve this high bandwidth requires that the exponent, g, of the power-law profile have a definite optimum value near $g = 2$ (for germanium dopant and an operating wavelength of 0.9 μm). A departure of only 0.05 from this optimum g value is sufficient to reduce the fiber bandwidth by more than one order of magnitude. Clearly the profiling technique must determine g to better than 0.05 if a meaningful correlation between



(a)

(b)

Fig. 1—Layer structure of MCVD fabricated preform as observed by: (a) diagnostrigram, and (b) index immersion.

fiber performance and index profile is to be obtained. In order to achieve this accuracy, $\Delta n(r)$ [the difference between the refractive index of the fiber core and its cladding value] must be measured with a precision of 1 part in 10^4.

Very slight local distortions of the refractive index profile from its optimum shape also decrease the fiber bandwidth markedly. A distortion of ten sinusoidal periods over the fiber radius with an rms amplitude of 0.6 percent of the maximum index difference reduces the bandwidth from 8000 MHz × km to about 200 MHz × km.[6] An rms distortion amplitude of only 0.15 percent for the same ripples reduces the bandwidth to about 800 MHz × km. The precision of the $\Delta n(r)$ measurement would again have to be about 1 part in 10^4 to detect even the 0.6 percent distortion. In addition, of course, the spatial resolution of the technique must be sufficient to resolve the perturbations.

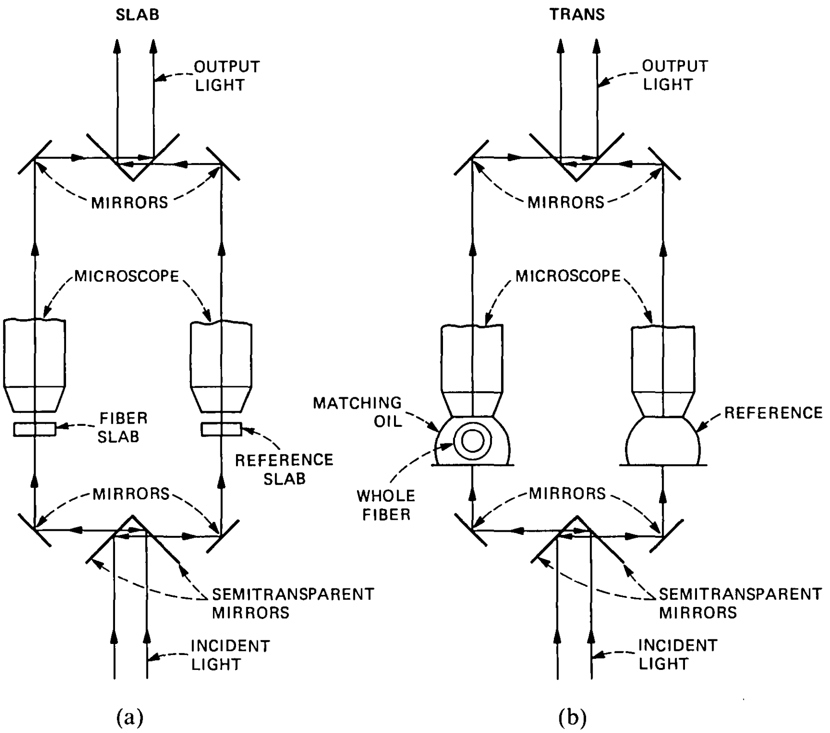


Fig. 2—Schematic diagram of (a) slab interferometric method, and (b) transverse interferometric method. The interference microscope used in both cases is identical.

II. PROFILE MEASUREMENT METHODS

The specific profiling methods used in this study are shown diagrammatically in Figs. 2 and 3. They are interference microscopy on SLAB samples (Fig. 2a); transverse interferometry on whole fiber samples (Fig. 2b); the focusing method (Fig. 3a); and the refracted near-field method (Fig. 3b). They are abbreviated by the terms SLAB, TRANS, FOCUS, and RNF, respectively. These methods will be discussed briefly; further details on their practical implementation can be found in the Appendix.

2.1 Slab interferometry

Interference microscopy on SLAB samples, utilizing the potential accuracy of interferometry was historically the first of these methods to be used,[7] and is generally accepted as the method to which newly developed techniques are compared. The SLAB sample is cut from an encapsulated fiber (or preform tip) and polished so that the faces are flat and parallel. The cutting and polishing procedures are both difficult and time-consuming. Special techniques are required to avoid

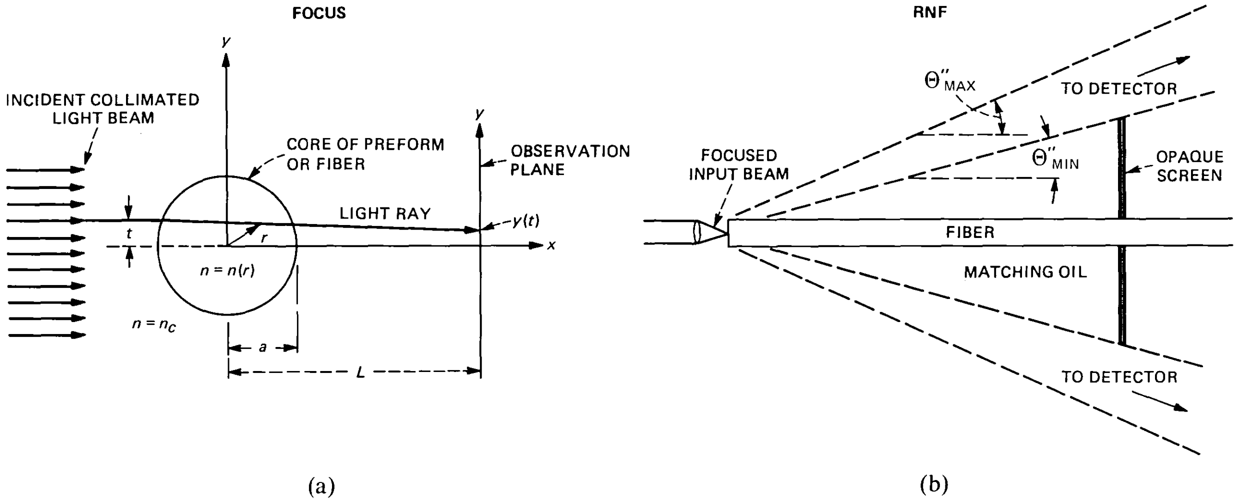


Fig. 3—Principle of operation of (a) focusing method, and (b) refracted near-field (RNF) technique.

composition-dependent thickness variations, which can lead to substantial errors.[8] It is also necessary for the sample to be thin enough so that rays traversing it are not bent and focused, producing curved wavefronts and, hence, erroneous results. Sample preparation requires about one day. This time can be reduced on a per-sample basis by processing several different fibers, which have been epoxied into one capillary tube, at the same time. It should also be noted that this procedure is not inherently nondestructive.

To observe the samples, of course, an interference microscope is required. Interference lens attachments to ordinary microscopes generally involve passing the light through the specimen twice, thus compounding possible errors. Best results are obtained with a single-pass Mach-Zehnder geometry, but microscope cost and availability are additional major considerations in adopting this method.

Relative index values accurate to about 2 parts in 10^4 can be realized routinely and by electronically processing the output of the microscope measurements relatively accurate to about 1 part in 10^5, as necessary, for example, in profile dispersion work, have been achieved.[9] Automatic computer processing of the output also serves to reduce analysis time. Spatial index resolution is somewhat limited in that it is not possible to combine maximum lateral resolution and exact phase measurements in a single instrument.[10]

2.2 Transverse interferometry

Sample preparation can be eliminated by using the transverse interferometric method (Fig. 2b). In this technique, a length of fiber is immersed in matching oil on the stage of the interference microscope and illuminated at right angles transverse to its axis. The matching oil removes the influence of the outer cladding boundary. The total optical path length of a light ray is expressed as an integral and the index distribution is obtained from the measured fringe shift by solving an integral equation.[11] Unlike the SLAB approach, in which the entire core is accessible, transverse interferometry assumes circular symmetry and, hence, geometry variations, which can adversely affect the profile, will not be detected unless special care is taken to make several measurements for different rotational positions of the fiber. By automating the measurement, index profiles of a fiber can be obtained within a few minutes after its manufacture. The accuracy of the method is about an order of magnitude less than that of the SLAB approach, and it is subject to a large error in the region near the fiber axis. On the other hand, this technique resolves detail in the fiber structure with higher resolution than the SLAB method, as will be seen later in the actual profiles.

2.3 Focusing method

The focusing method[12] (Fig. 3a) is similar to transverse interferometry in that it is also nondestructive and uses transverse illumination. Otherwise they are very different. The focusing method does not require an interference microscope or rely in any way on interferometry. Moreover, the technique is readily applicable with high accuracy to fiber preforms.[3]

In this method, the fiber, observed with a microscope and the preform with a camera lens, are immersed in index matching fluid. The core, acting as a cylindrical lens, focuses the light, whose power density distribution in the observation plane is detected by a video camera. After digitizing the power density a computer calculates the index profile by solving an integral equation. The profiles so obtained from circularly symmetric cores are comparable in accuracy and resolution to those produced by interference microscopy of SLAB samples.[13]

Extreme experimental care, however, is involved in the focusing method since it measures absolute light intensities. The optics, matching oil, and fiber (or preform) itself must be very clean; the video detector must be linearized and the incident light intensity must be uniform, either intrinsically or through a calibration procedure.

2.4 Ray tracing

A related method to measure the profile in preforms is ray tracing.[14] Further details of this technique are also given in the Appendix. This method involves scanning a focused laser beam perpendicular to the axis of the index-matched preform and recording the exit angle of the beam as a function of distance from the axis. The profile of the preform is then reconstructed by taking the inverse Abel transform of the deflection function. Indeed, the mathematics of this and the focusing method are nearly identical[12] and lead to very similar profiles, if an equal number of data points are measured and processed.

2.5 Refracted near-field

The refracted near-field (RNF) method[15] (Fig. 3b) relies on the power escaping sideways from the core into the cladding to determine the refractive index profile of the fiber. The fiber, immersed in a matching oil whose index of refraction is greater than that of the cladding, is passed through a small hole in an opaque disc. Part of the light focused into the fiber is guided while the rest appears outside of the fiber as a hollow cone. If all of the leaky modes, contained in the inner part of this cone are blocked by the disc, then the light passed varies linearly with the index of refraction of the fiber at the point at which the incident light is focused. Thus, by scanning the incident light across

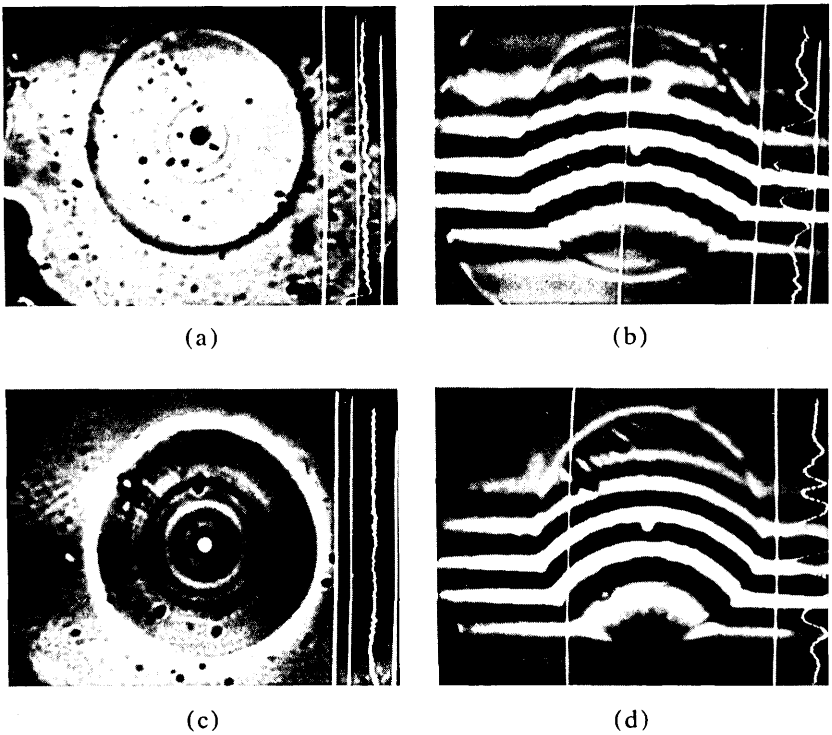


Fig. 4—Two fiber-SLAB samples observed by (a,c) ordinary microscopy, and by (b,d) interference microscopy.

the end of the fiber, the profile is obtained directly from the output of a detector that collects the light passing the disc.[16] This method, as will be seen in the profiles, has excellent spatial resolution and also possesses the ability to analyze noncircular cores. The precision of the index measurement is estimated in one recent embodiment[16] to be about 4×10^{-5}.

III. MEASUREMENT RESULTS

The fiber samples used in this study were specifically chosen to possess a variety of features, which would severely test the limits of the different profiling methods.

A SLAB sample of the first fiber is shown in Figs. 4a and 4b as observed with ordinary and interference microscopy, respectively. The fiber is seen to possess severe perturbation in layer structure throughout the core with marked variations especially near the center. The index profiles as obtained by the SLAB, FOCUS, and TRANS methods are shown in Fig. 5 by the solid, dotted, and dashed curves, respectively. The index and core radius values, which are in very good agreement, are those obtained by each of the measurements and no scaling was employed. Generally the profile shapes are similar, with the TRANS profile exhibiting more detailed structure of the perturbations. Slight differences are accounted for by the assumption of circular symmetry in the FOCUS and TRANS cases and the lack of such an assumption for the SLAB.

A comparison of the RNF profile (solid curve) and the SLAB profile (broken curve) for this fiber is shown in Fig. 6. The SLAB, FOCUS, and TRANS profiles were all measured by the author at Bell Laboratories, Crawford Hill. The RNF profiles were provided by Jeff Saunders at Bell Laboratories, Atlanta.[16] The resolution of structure in the RNF profile is striking in comparison to the SLAB, which appears as a near average through the RNF results. The scale here and in subsequent profiles has been eliminated for clarity and ease of comparison. The RNF profile also shows a steeply rising region, and an index depression at the core-cladding interface, features not well resolved by the SLAB. The SLAB measurement is also not able to resolve the central depression due to its steep gradient.

A comparison of the RNF profile (solid curve) and the TRANS profile (broken curve) is shown in Fig. 7. The superior resolution of the TRANS

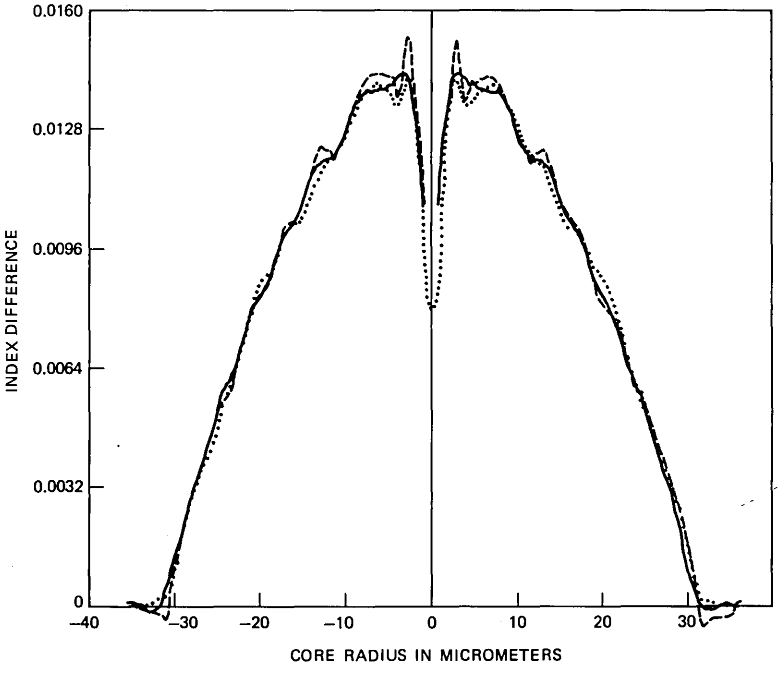


Fig. 5—Profiles of fiber shown in Figs. 4a and 4b (fiber no. 1) as obtained by SLAB (solid), FOCUS (dotted), and TRANS (broken) techniques.

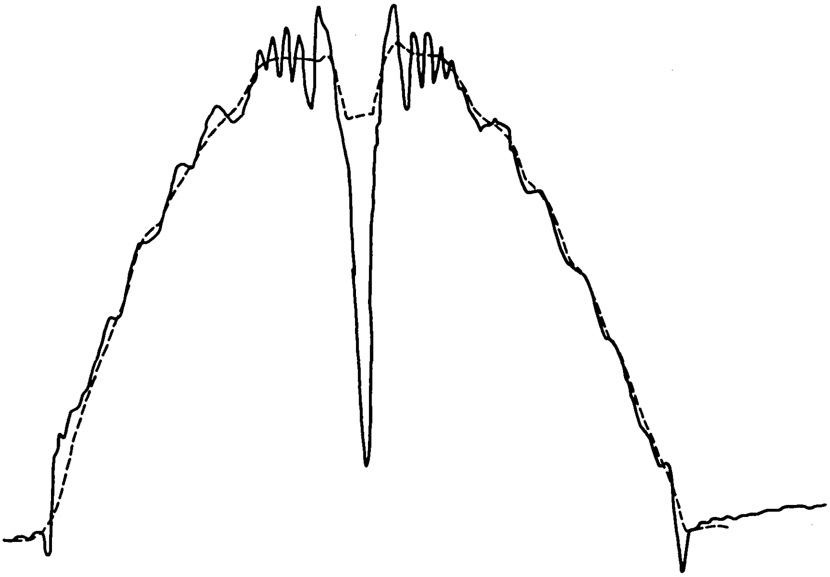


Fig. 6—Comparison of SLAB (broken) and RNF (solid) profiles for fiber no. 1.

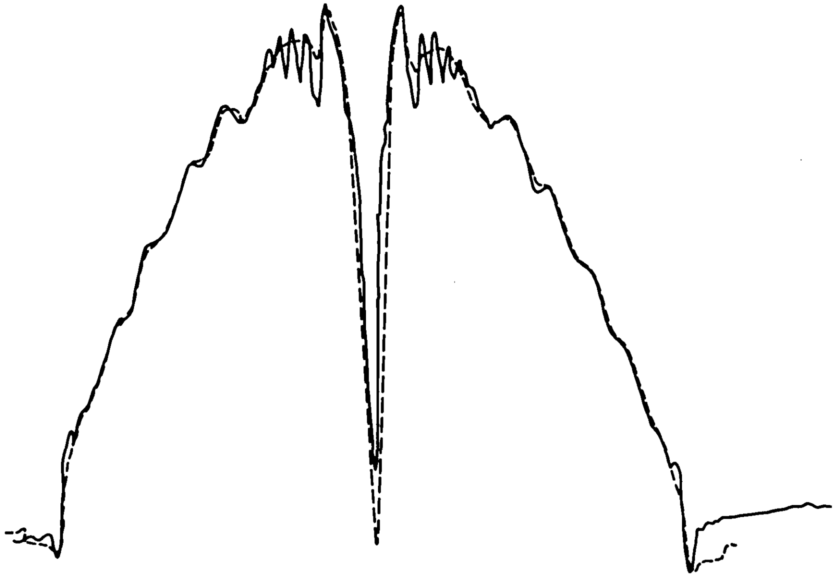


Fig. 7—Comparison of RNF (solid) and TRANS (broken) profiles for fiber no. 1.

measurement to that of the previously shown SLAB is clearly seen in that now the curves are practically identical, except for the region near the center where RNF shows greater detail.

The comparison of RNF (solid curve) and FOCUS (broken curve)

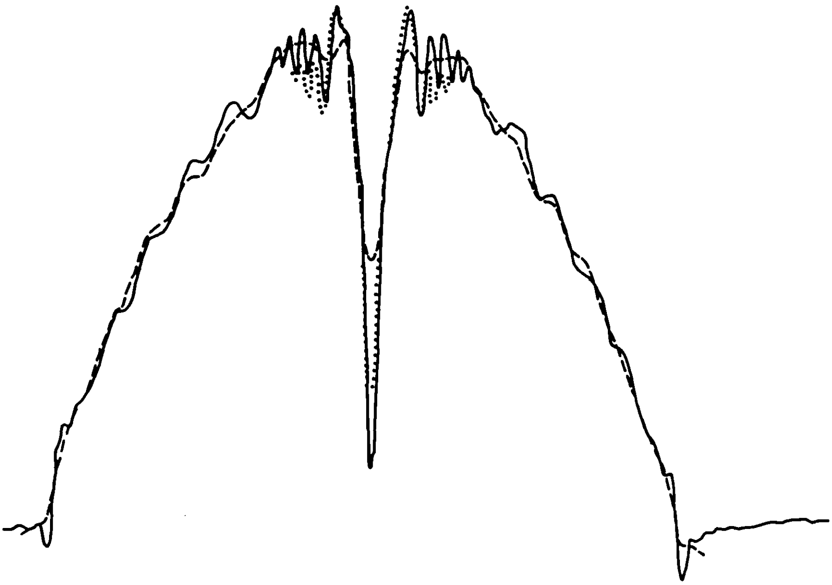


Fig. 8—Comparison of RNF (solid) and FOCUS (broken) profiles for fiber no. 1. The dotted curve is a portion of a focused profile obtained by focusing within the core.

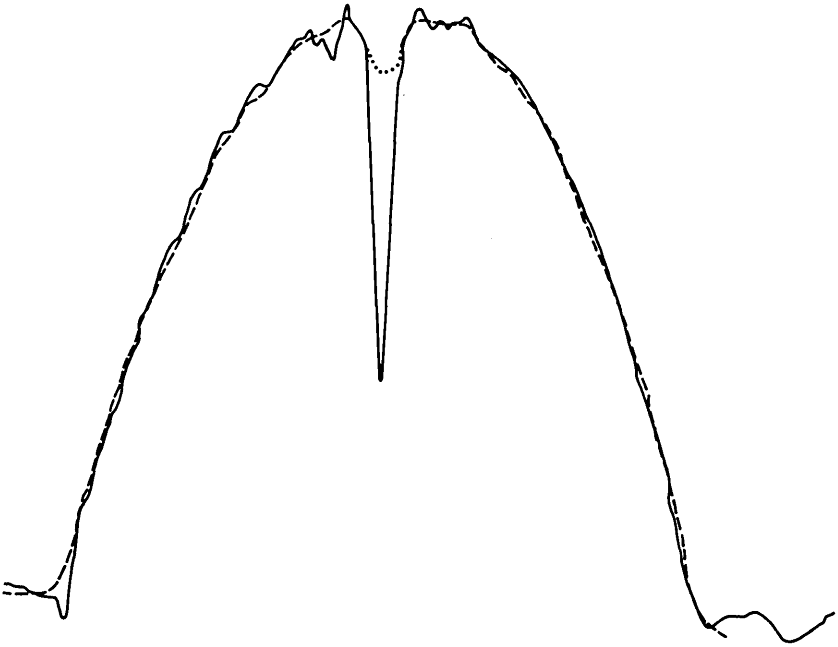


Fig. 9—Comparison of RNF (solid) and SLAB (broken) profiles for fiber shown in Figs. 4c and 4d (fiber no. 2).

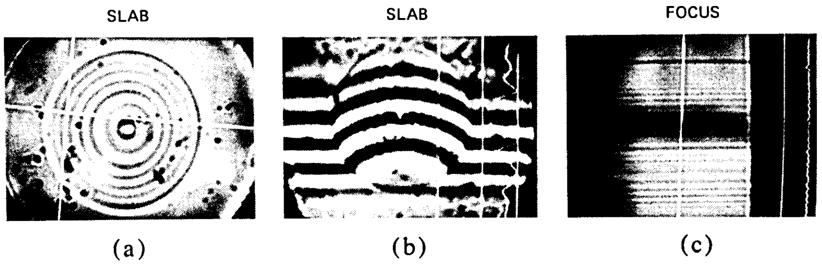


Fig. 10—Fiber no. 3 as observed by (a) ordinary microscopy, (b) slab interference microscopy, and (c) the focusing method.

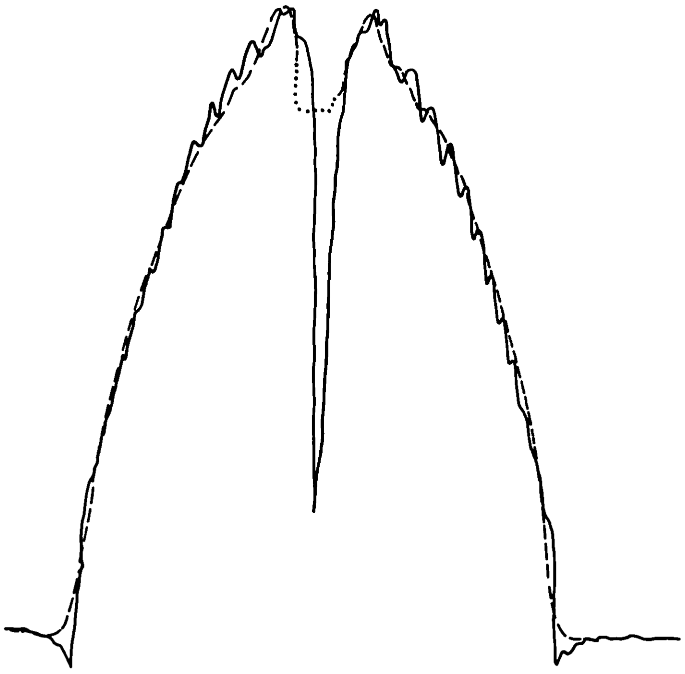


Fig. 11—Comparison of RNF (solid) and SLAB (broken) profiles for fiber no. 3.

shown in Fig. 8 is similar to the RNF versus SLAB results. A second FOCUS profile, a portion of which is shown by the dotted curve, was obtained by focusing within the core to obtain greater resolution. Indeed the resolution in this region is now comparable to RNF but the remainder of the profile (not shown) is distorted because of the violation of the focusing condition. Focusing within the core actually satisfies the focusing condition locally for those rapid variations that tend to focus the incident rays much more steeply than the remainder of the core.

A second SLAB sample with somewhat smaller index variations is shown in Figs. 4c and 4d, again by ordinary and interference microscopy, respectively. A comparison of the RNF (solid curve) and SLAB (broken curve) profiles is presented in Fig. 9. Because of the smaller index fluctuations, the curves agree quite well but, again, the high resolution of RNF is apparent.

A third SLAB sample of a fiber fabricated with several (~10) discrete changes in dopant concentration as a function of core radius is shown in Figs. 10a and 10b by normal and interference microscopy, respectively. Figure 10c shows the whole fiber sample as observed by the focusing method.

A comparison of the RNF (solid curve) and SLAB (broken curve) profiles seen in Fig. 11 shows good agreement as far as the general profile shape is concerned, but the SLAB curve is, again, almost an average through the well-resolved layer structure of RNF. The FOCUS profile is similar to the SLAB, except when it is obtained by focusing within the core in which case specific local features can be brought out, at the expense of an overall distortion, with resolution comparable to RNF.

The TRANS profile of this fiber, shown by the broken curve in Fig. 12, is in excellent agreement with the RNF profile (solid curve). The resolution of the layers, the widths of which are somewhat larger than

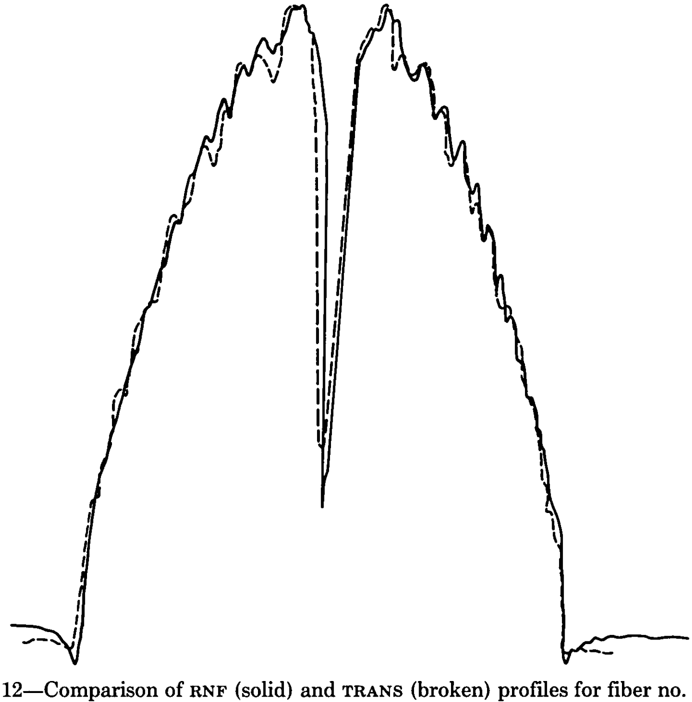
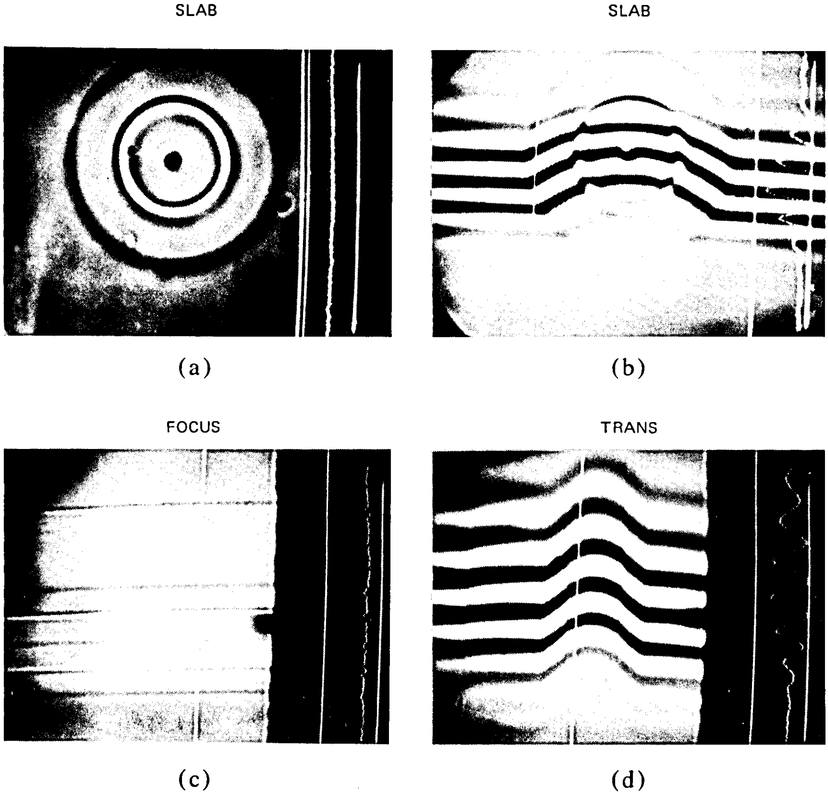


Fig. 12—Comparison of RNF (solid) and TRANS (broken) profiles for fiber no. 3. Note the comparable resolution.



SLAB SLAB

(a) (b)

FOCUS TRANS

(c) (d)

Fig. 13—Fiber no. 4, possessing a strong index perturbation, as observed by the various measurement methods.

in the first sample, is equally good. The index depression and steep rise in the index distribution are also the same in both.

A fourth sample having a major index perturbation at 50 percent of the radius is shown in Figs. 13a and 13d by ordinary microscopy, SLAB interferometry, focusing method, and transverse interferometry, respectively. The perturbation is clearly seen in each method of observation. A comparison of the SLAB (broken curve) and RNF profile (solid curve) is shown in Fig. 14. The profile shapes are similar but the index distortion is more prominent in the RNF measurement. The TRANS profile, shown as the broken curve in Fig. 15, on the other hand, shows the perturbation as clearly resolved as RNF (solid curve). The TRANS profile also displays the fine index variations, as does the RNF result, which lie closer to the center of the core. These fluctuations are absent in the SLAB and also in the FOCUS profiles both of which are very similar.

Finally, a fiber with a relatively perturbation-free index distribution

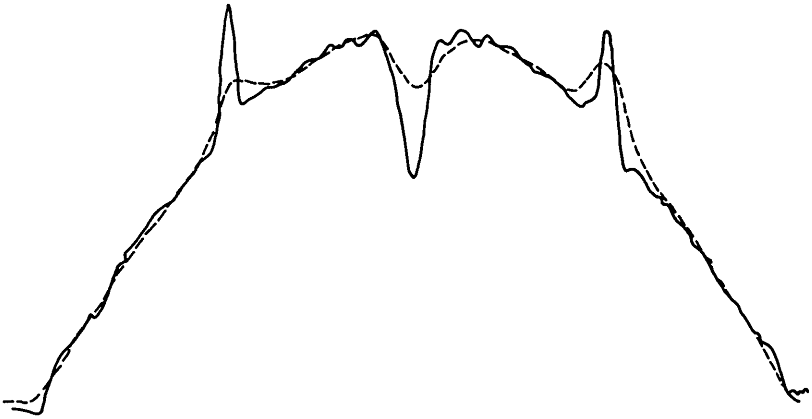


Fig. 14—Comparison of RNF (solid) and SLAB (broken) profiles of fiber no. 4.

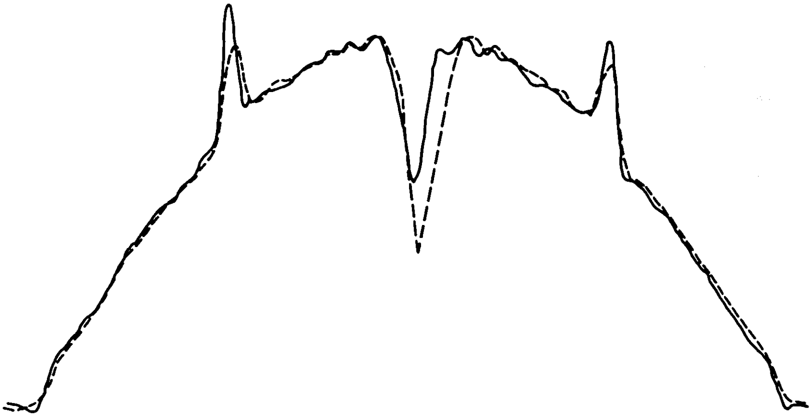


Fig. 15—Comparison of RNF (solid) and TRANS (broken) profiles for fiber no. 4.

was studied. The fiber as viewed by the SLAB, TRANS, and FOCUS techniques is shown in Fig. 16. To be noted in particular is the very uniform appearance (except for dirt specks) of all the samples indicating a lack of index distortions. The profiles obtained by the different methods, as might be imagined, are all very similar. The RNF (solid curve) and SLAB profiles (broken curve) are shown in Fig. 17 and the RNF (solid), FOCUS (dotted), and TRANS (broken) profiles are seen in Fig. 18. The only differences in the index distributions is a steep initial rise in the profile and some very slight ripple seen in the RNF result near the core center. This particular fiber also had a strong asymmetry in the index distribution near the center. This is not apparent in the FOCUS and TRANS results since they depend upon the assumption of circular symmetry.

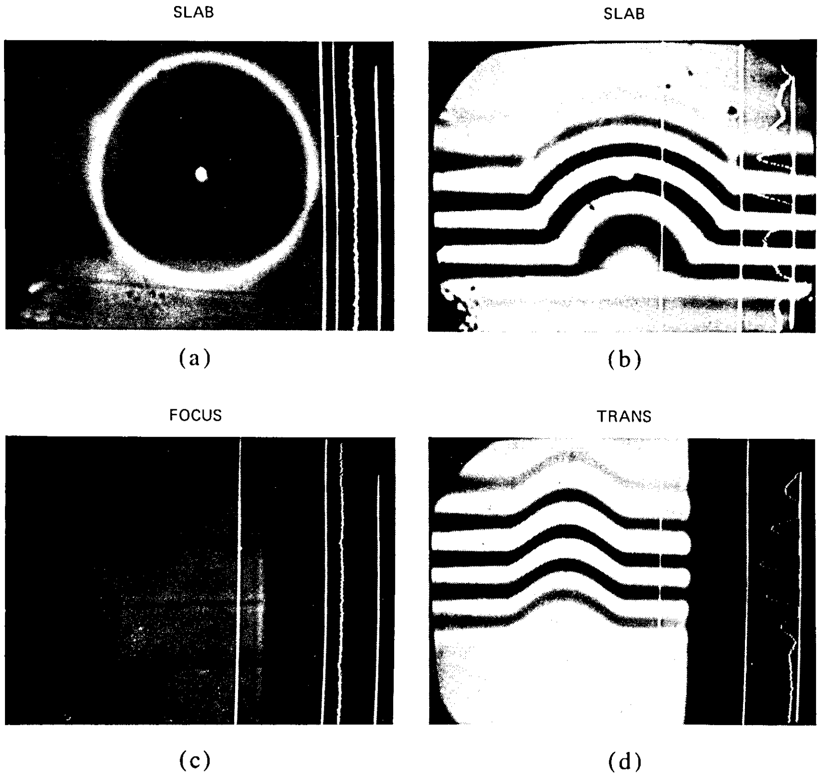


Fig. 16—Fiber no. 5 possessing a relatively perturbation-free profile as observed by the various measurement methods.

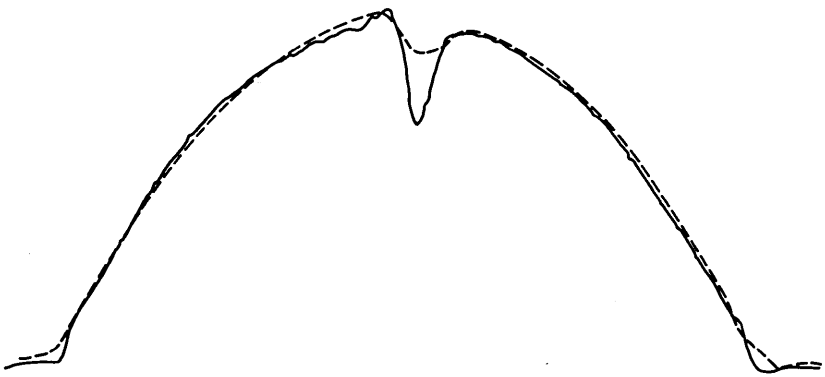


Fig. 17—Comparison of RNF (solid) and SLAB (broken) profile for fiber no. 5.

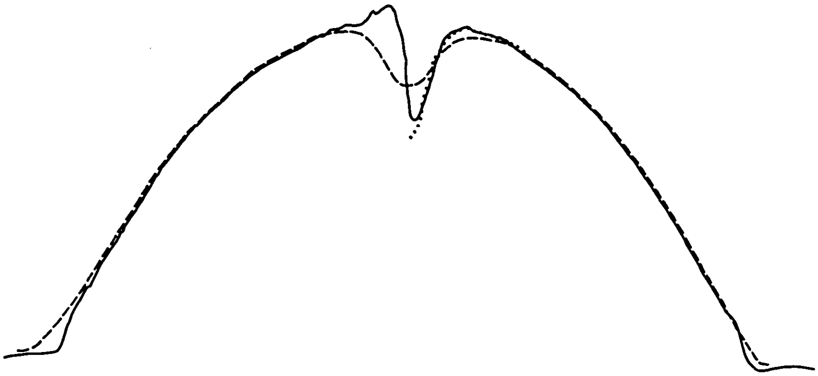


Fig. 18—Profiles of fiber no. 5 as measured by RNF (solid), FOCUS (dotted), and TRANS (broken) techniques.

An important question is the relationship of profiles measured in the fibers to the actual profile existing in the corresponding preforms. A knowledge of this relationship would shed additional light on the applicability of the various methods and give confidence that structural features observed in the fiber profiles are not measurement artifacts. Of particular interest is the steep initial rise in the index distribution of the last fiber as seen by RNF but not clearly defined with the other methods. Is this feature in the preform or not?

To answer this question the preform corresponding to the last sample was profiled by the two methods previously described, the focusing method as applied to preforms[3] and the ray-tracing technique as used at Western Electric's Engineering Research Center.[14]

The preform profiles are shown in Fig. 19; the solid curve is obtained by the ray-tracing technique and the broken curve by the focusing method. The profiles can barely be distinguished, except in the region near the center where the previously mentioned index perturbation makes the profile shape orientation dependent, and in the resolution of the finer deposition layers, which the ray-tracing method achieves by processing about ten times as many data points as the focusing method.

A comparison of the focused preform profile and the RNF fiber profile is shown in Fig. 20. Scaling of the radial coordinate, of course, was performed but not of the index values. The agreement of the profiles is excellent. It is seen that the initial steep rise which appears in the RNF profile is indeed present in the preform. It does, however, only appear as a single step in the RNF result, whereas in the preform it has a fine double-step structure. The chosen resolution of the focused preform profile is, thus, seen to correspond to the actual fiber profile. The greater detail included in the ray-tracing technique is absent even

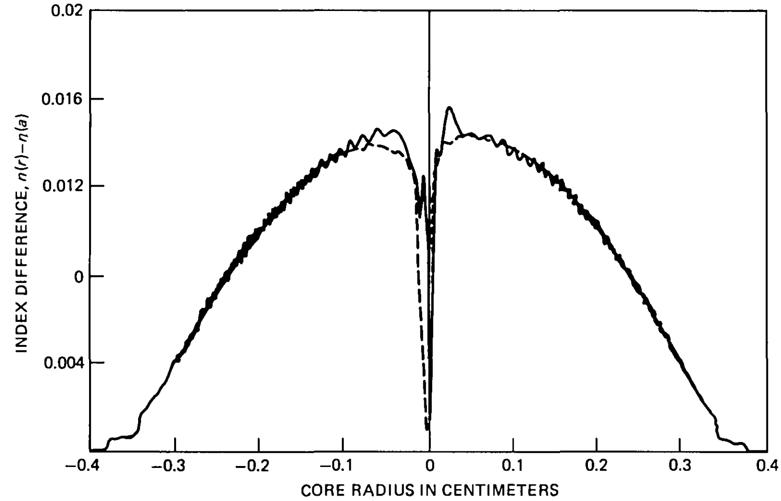


Fig. 19—Profiles of the preform stub from which fiber no. 5 was pulled as measured by the ray-tracing technique (solid curve) and the focusing method (broken). The two halves of each profile were obtained from measurements at two orientations.

from the RNF-profiled fiber, because of its small-scale structure. In general, a very high degree of correspondence between fiber and preform profiles exists.

The various fiber profiles were further compared by taking the rms deviation of their differences as a function of radial position. This is the severest possible comparison in that absolute point-by-point values are looked at. Thus, a slight shift of a feature could result in a sizable difference. A second less restrictive comparison was also made by computing the best-fit g values to the various profiles. This emphasizes the profile shape at the expense of the location and existence of small perturbations.

It was found that the various profiling methods give just about identical results for relatively smooth profiles. The rms deviations of the profiles of the last fiber (shown in Figs. 16–20), were all less than one percent, excluding the index depression in the center. The central depression gives rise to a few percent difference on its own because of the different ways it is resolved by the various methods. The best fit g curves were all within 0.05 of each other. The RNF profile has a g value of 2.024 and that of the preform (as measured by the focusing method) a value of 1.990, a difference of less than 0.035.

As expected, the rms differences between the profiles increase for those possessing rapid variations. These differences for the first four fiber samples shown amounted to three to five percent, again excluding the index depression. The g values of the respective profiles were all

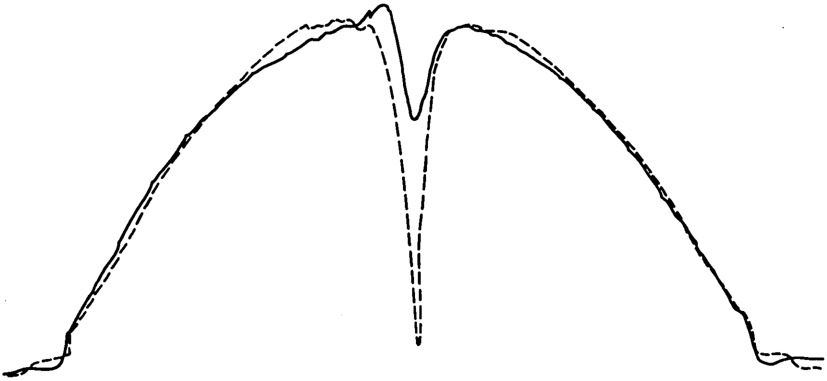


Fig. 20—Comparison of preform profile (broken) and corresponding RNF fiber profile (solid) for fiber no. 5.

within a few percent, as would be expected from the similarities of the curves.

We conclude then, that as long as the profile is smooth all of the measurement methods are equivalent. For profiles containing rapid perturbations, the RNF and TRANS methods are about comparable with the former doing better in resolving very rapid variations. The RNF method also has the advantage of not requiring an interference microscope nor elaborate computer processing of the data. On the other hand, it does require a separate calibration procedure and reasonable care with cleanliness of the optics.

The fact that the rms deviations are less than one percent and the best-fit g values are within 0.05 for relatively perturbation free profiles lends confidence to the ability to make valid bandwidth predictions for current fibers of this type based on the various measured profiles.[17] Deviations of this magnitude, while reducing the bandwidth by about one order of magnitude, still result in bandwidths close to 1 GHz. As profiles get even smoother these deviations will presumably decrease and allow for even better predictions. Meanwhile, improvements in the accuracy of the profiling methods themselves, can also serve to reduce these interprofile deviations and lead to even better agreement. An improvement in accuracy of about a factor of 5 is required to meaningfully measure deviations of the ideal profile. On the other hand, one then enters the realm of theoretical uncertainty as to what the ideal ideally is. The current state-of-the-art of index measurements should then be able to go a long way in providing the feedback necessary to improve current profiles before their limitations are indeed felt.

While we have presented four current profiling techniques, there are, of course, other existing methods each subject in use to their own set of trade-offs. New methods both for fibers and preforms are reported regularly and undoubtedly as they prove their value will find use in the important task of index profiling.

IV. ACKNOWLEDGMENTS

The author thanks J. Saunders for providing the RNF profiles, L. Watkins for supplying the ray-tracing profile of the preform, and D. Marcuse for computational analysis.

APPENDIX

Implementation of Profile Measurement Methods

Preform diagnostigrams

Preform diagnostigrams provide a sensitive, nondestructive, and noncontacting means of obtaining structural information. This information includes a measurement of core size and core eccentricity, core cladding interface structure, individual deposition layer structure and variations, imperfections within the core and the cladding, and the presence of an axial refractive-index depression.

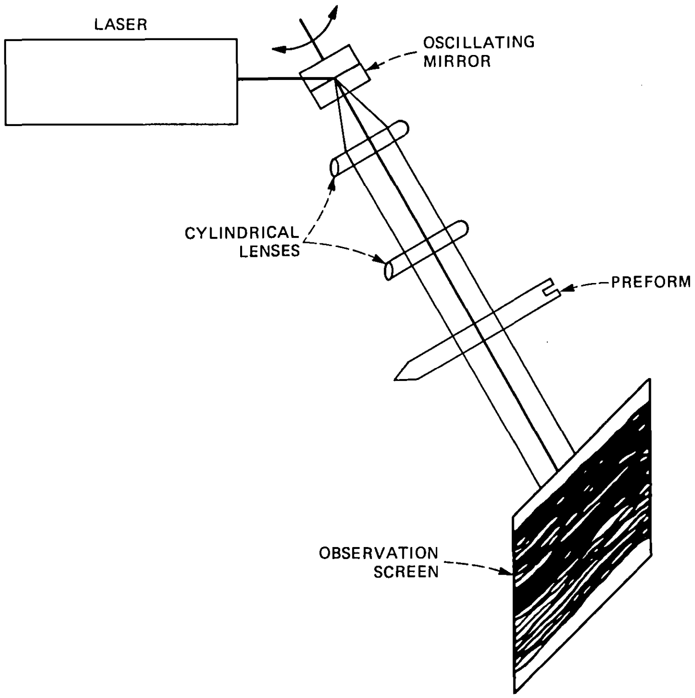


Fig. 21—Arrangement to produce preform diagnostigrams.

A diagnostigram is produced by the arrangement shown in Fig. 21. Light from a He-Ne laser is incident upon an oscillating mirror that serves to transform it into a line. The line of light is expanded and collimated by two cylindrical lenses and traverses the preform. The beam is about equal in width to the diameter of the preform, and its length can be varied as desired by adjusting the amplitude of the oscillating mirror. Typically a length of 15 cm is used. An arrangement of cylindrical lenses replacing the oscillating mirror can also be utilized. The light traversing the preform is then incident upon either an observation screen or photographic film.

The pattern, as shown in Fig. 1a (of the text), consists of bright and dark lines. The width of the bright lines represents the actual geometric width of discontinuous index steps in the core, while the width of the dark lines represents the amount Δn of refractive-index discontinuity. The discontinuities arise during the deposition process and represent distinct deposition layers. There exists a one-to-one correspondence between the observed lines and the deposited layers. Further details can be found in Ref. 2.

Slab and transverse interferometry

In slab interferometry the fiber sample is placed in one arm of the interference microscope and a homogeneous reference SLAB with refractive index n_2 is placed in the reference arm (Fig. 2a). An arrangement of practical implementation is shown in Fig. 22. Figure 23 displays the fringe shifts observed in a graded-index sample, the shift S of a fringe depending on its position in the fiber core, $S = S(r)$. The difference between the refractive indices of core and cladding can be expressed in terms of this fringe shift $S(r)$, the uniform fringe spacing in the cladding D, the vacuum wavelength of light λ, and the SLAB thickness t as

$$n(r) - n_2 = \frac{\lambda S(r)}{Dt}. \tag{1}$$

To measure the fringe shift a video camera looks into the interference microscope and sends its electrical output signal to a video digitizer. The 8-bit digitizer is computer controlled and addresses specific, preselected points in the video field. A video monitor and a plotter for recording the processed information—the index profile, is also included. The computer directs the vertical sample line seen in Fig. 23 to collect information on either side of the core (along lines A-B, and C-D), which is then used to determine the fringe spacing and to compensate for a tilt of the entire fringe pattern. The computer then advances the sample line in small increments, moving it through the core region, measuring the displacement of the fringe that goes through

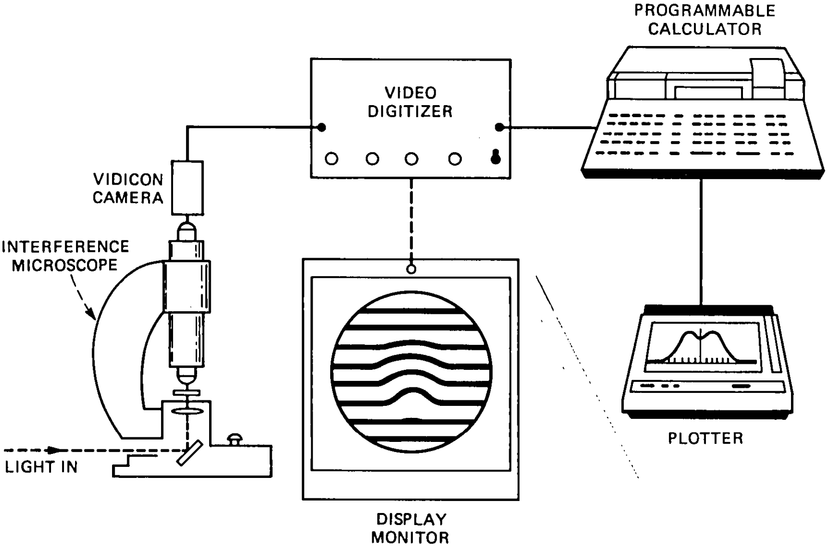


Fig. 22—Arrangement for automatic refractive index profiling of SLAB samples.

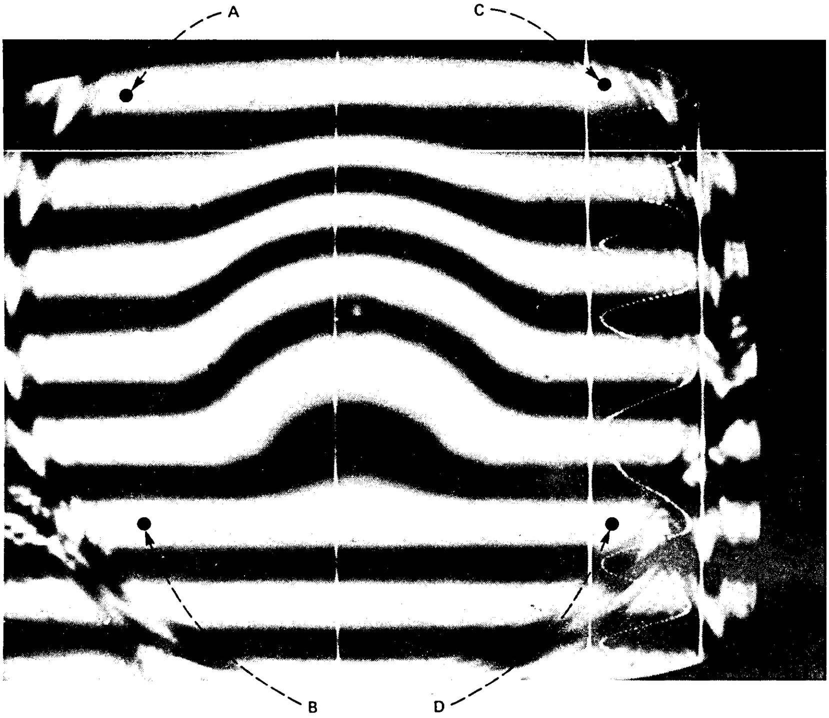


Fig. 23—Fringes observed in a graded-index SLAB sample.

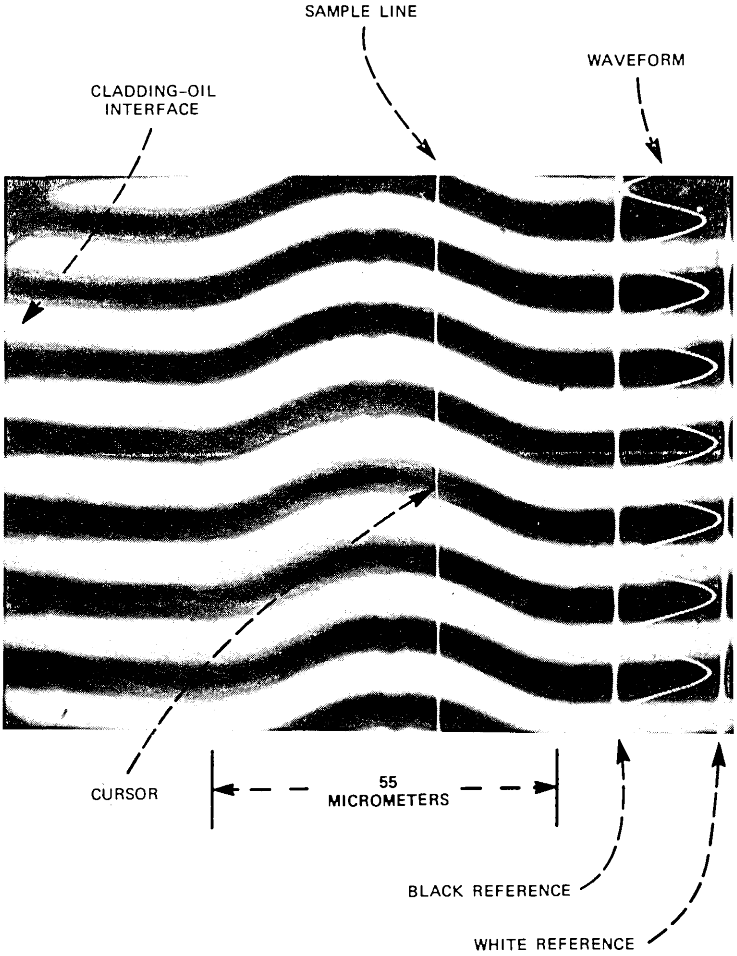


Fig. 24—Fringes observed in a graded index whole fiber sample observed transversely.

the core center. The computer determines the fringe positions by counting and searching for the minimum light level whose location it pinpoints by least mean square fitting of a parabola using a number of points in the vicinity of the minimum.

The fringe displacement is recorded as a function of the radial coordinate r measured from the core center and the resulting function $S(r)$ is used to compute $n(r) - n_2$ according to eq. (1). The index distribution is then sent to the plotter.

In transverse interferometry the fiber, after being stripped of any jacket, is again placed in one arm of the interference microscope (Fig. 2b in text). The fiber is covered with a drop of matching oil into which the microscope objective is dipped. The reference branch contains only a drop of matching oil. Each light ray incident upon the sample now passes through regions of varying refractive index and the total path length must be expressed as an integral. The refractive index difference between core and cladding is given by

$$n(r) - n_2 = \frac{\lambda}{\pi D} \int_r^\infty \left| \frac{dS(\rho)}{d\rho} \right| \frac{d\rho}{\sqrt{\rho^2 - r^2}} \tag{2}$$

in which ρ is the radial coordinate and rotational symmetry is assumed.

The output field of the interference microscope now appears as in Fig. 24, and the fringe shift is measured with the computer-controlled video digitizer system just described. Processing of the fringe shift information requires that a numerical differentiation is performed first, followed by the numerical integration indicated by eq. (2).

The focusing method

The focusing method is shown in application to fibers and preforms in Fig. 25. The technique uses incoherent filtered light in transverse illumination. The fiber, observed with a microscope, and the preform, observed with a camera lens, are immersed in index matching fluid to avoid the deleterious influence of the outer cladding boundary. The core, acting as a cylindrical lens, focuses the light whose power density distribution in the observation plane is detected by a video camera. The observation plane is defined by the object plane on which the camera is focused. This plane must not be inside the fiber core, and it must not be placed so far away that rays have already crossed over after leaving the core. Good results are obtained when the observation plane is placed just outside of the core.

The image of a preform seen by the camera is shown in Fig. 1b of the text. A monitor display of a fiber is shown in Fig. 26.

The refractive index distribution is obtained by solving the integral equation

$$n(r) - n_2 = \frac{n_2}{\pi L} \int_r^\infty \frac{t - y(t)}{\sqrt{t^2 - r^2}} \, dt. \tag{3}$$

The various parameters are defined in Fig. 3a of the text.

The function $y(t)$ is obtained from a measurement of the light power density distribution in the observation plane. This distribution is collected along the sample line and digitized under computer control as described previously. The computer also solves the integral equation and plots the resultant index profile.

Refracted near-field method

In this technique[15] (shown in Fig. 3b of the text) a light beam is focused on a spot at a distance r from the fiber axis with a convergence

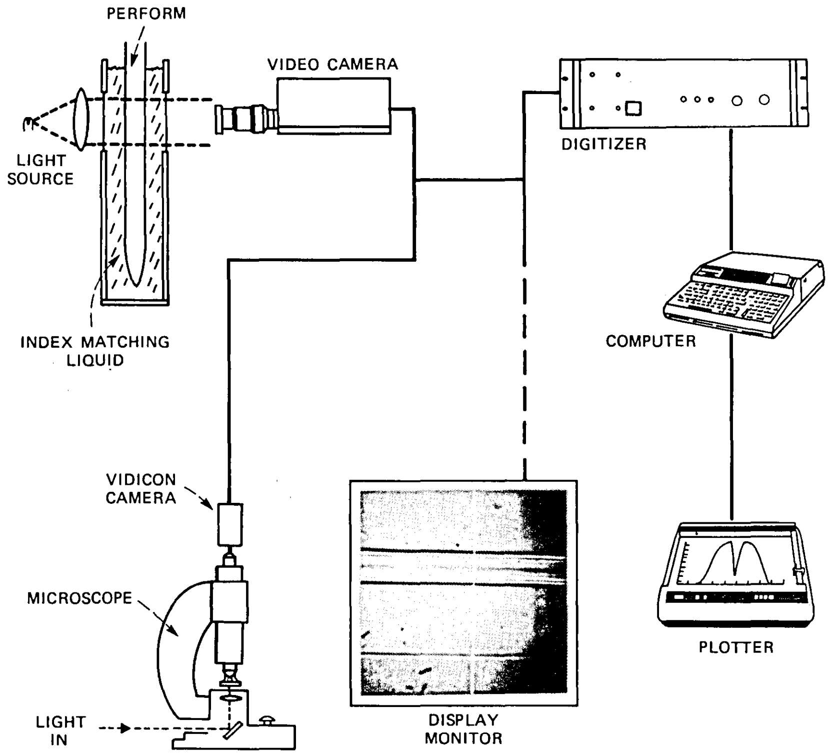


Fig. 25—Arrangement for automatic refractive index profiling of fibers and preforms by the focusing method.

angle that is larger than the acceptance angle of the fiber core. The light escaping from the core is partly contributed by power leakage from leaky modes. This part of the radiated power is blocked by a circular aperture which prevents light leaving below a minimum angle $\theta''_{\min}$ from reaching the detector. The refractive index difference between the core and cladding, which is obtained from the light passed by the aperture, is given by

$$n(r) - n_2 = n_2\cos\theta''_{\min}[\cos''_{\min} - \cos\theta''_{\max}]\frac{P(a) - P(r)}{P(a)}. \qquad (4)$$

In this expression θ'' refers to the input angle, $P(r)$ is the light power reaching the detector, as a function of the position of the input beam and $P(a)$ is obtained from the $P(r)$ curve as the light power detected when the input beam is focused into the cladding.

The experimental apparatus used in the implementation of this method by J. Saunders is shown in Fig. 27.[16] Light from a 5-mW He-Ne laser passes through a quarter wave plate and is focused onto a 50-

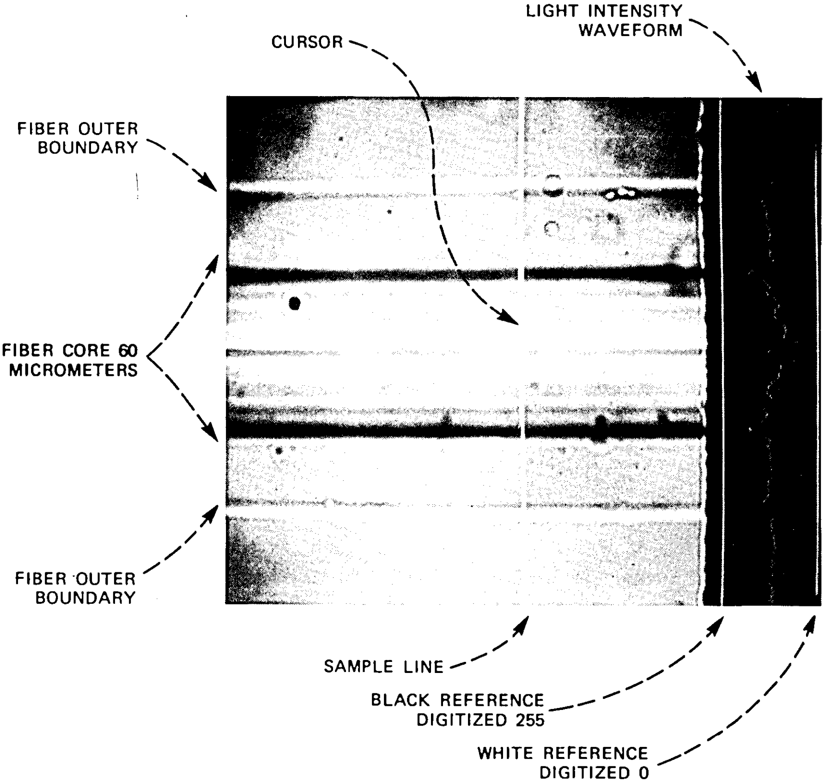

LIGHT INTENSITY WAVEFORM
CURSOR
FIBER OUTER BOUNDARY
FIBER CORE 60 MICROMETERS
FIBER OUTER BOUNDARY
SAMPLE LINE
BLACK REFERENCE DIGITIZED 255
WHITE REFERENCE DIGITIZED 0

Fig. 26—Graded-index fiber as observed by the focusing method.

μm pinhole. The light from the pinhole is then focused by a 20x microscope objective onto the end face of the fiber which is held in a moveable cell containing index immersion liquid.

The disc that blocks the leaky rays is 1.3 cm in diameter and is supported by means of three fibers. The light passing the disc is directed by lenses to a large area detector whose output provides the profile. The microscope and TV camera provide a magnified view of the fiber core for alignment and monitoring purposes.

Ray-tracing method

The practical implementation of this method by L. S. Watkins, is shown in Fig. 28.[14] A narrow beam from a He-Ne laser is reflected off a rotatable mirror and is focused by a lens through the index-matched preform. Rotating the mirror moves the ~20 μm beam across the preform.

The deflected beam is collected by a lens and focused onto a linear position sensor at its back focal plane. The output of the sensor is analyzed to give a voltage proportional to the deflection angle, which is then computer-processed in a similar manner to the focusing method to give the profile.

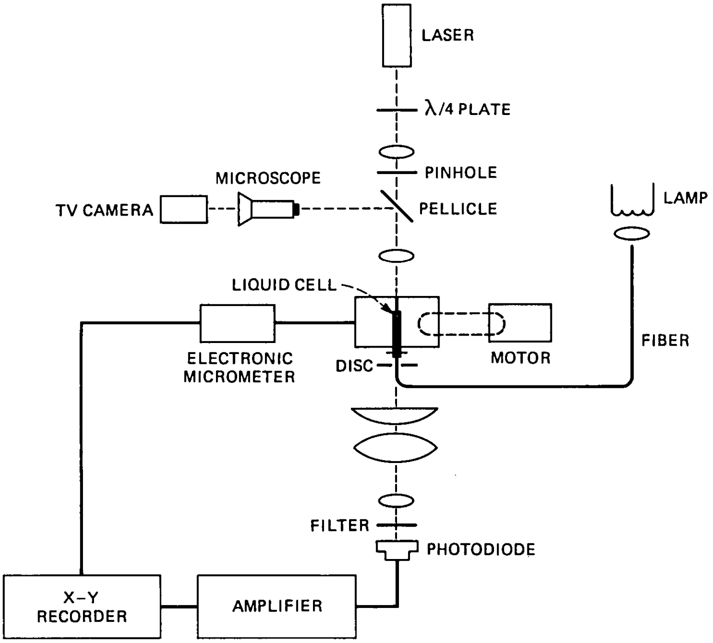


Fig. 27—Experimental apparatus for implementation of the RNF method.

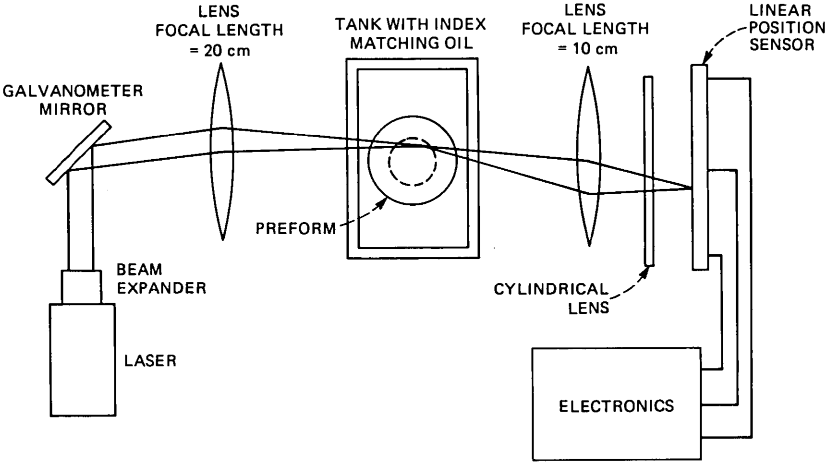


Fig. 28—Arrangement for profiling of preforms by the ray tracing method.

REFERENCES

1. J. B. MacChesney, P. B. O'Connor, and H. M. Presby, "A New Technique for the Preparation of Low Loss and Graded-Index Optical Fibers," Proc. IEEE, *62* (September 1974), pp. 1280-1.
2. H. M. Presby and D. Marcuse, "Optical Fiber Preform Diagnostics," Appl. Opt., *18* (1979), pp. 23-30.
3. H. M. Presby and D. Marcuse, "Preform Index Profiling (PIP)," Appl. Opt., *18* (March 1979), 671-77.
4. H. M. Presby, et al., "Material Structure of Germanium-Doped Optical Fibers and Preforms," B.S.T.J., *54,* No. 10 (December 1975), pp. 1681-92.
5. S. Suzuki et al., "Transmission Characteristics of Graded-Index Fiber," Technical Digest, 1977 Int. Conf. Integrated Optics and Optical Fiber Commun. (IOOC) (1977) p. 459.
6. D. Marcuse and H. M. Presby, "Effects of Profile Deformations on Fiber Bandwidth," Appl. Opt., *18* (1979) p. 3758.
7. C. A. Burrus and R. D. Standley, "Viewing Refractive-Index Profiles and Small-Scale Inhomogeneities in Glass Optical Fibers: Some Techniques," Appl. Opt., *13* (1974) p. 2365.
8. J. Stone and R. M. Derosier, "Elimination of Errors Due to Sample Polishing in Refractive-Index Profile Measurements by Interferometry," Rev. Sci. Instrum., *47* (1976), p. 885.
9. H. M. Presby and H. W. Astle, "Optical Fiber Index Profiling by Video Analysis of Interference Fringes," Rev. Sci. Instrum., *49* (1978), pp. 339-44.
10. E. Ingelstam and I. P. Johansson, "Correction Due to Aperture in Transmission Interference Microscopes," J. Sci. Instrum., *35* (1958), p. 15.
11. H. M. Presby, et al., "Rapid Automatic Index Profiling of Whole-Fiber Samples: Part II," B.S.T.J., *58,* No. 4 (April 1979), pp. 883-902.
12. D. Marcuse, "Refractive Index Determination by the Focusing Method," Appl. Opt. *18* (1979), pp. 9-13.
13. D. Marcuse and H. M. Presby, "Focusing Method for Nondestructive Measurement of Optical Fiber Index Profiles," Appl. Opt., *18* (1979), pp. 14-22.
14. L. S. Watkins, "Laser Beam Refraction Transversely Through a Graded-Index Preform to Determine Refractive Index Ratio and Gradient Profile," Appl. Opt., *18* (1979), 2214-22.
15. W. J. Stewart, "A New Technique for Measuring the Refractive Index of Graded Optical Fibers," Technical Digest, 1977 Int. Conf. Integrated Optics and Optical Fiber Commun. (IOOC) Tokyo, Japan, July 18-20, 1977.
16. M. J. Saunders, "Optical Fiber Profiles Using the Refracted Near Field Technique: A Comparison With Interferometry," Technical Digest Supplement, Symp. Optical Fiber Measurements, Boulder, Colorado, October 28-29, 1980.
17. H. M. Presby, D. Marcuse, and L. G. Cohen, "Calculation of Bandwidth from Index Profiles of Optical Fibers. Part 2: Experiment," Appl. Opt., *18* (1979), pp. 3249-55.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 60, No. 7, September 1981
Printed in U.S.A.

Statistical Behavior of Crosstalk Power Sum With Dominant Components

By S. H. LIN

(Manuscript received February 4, 1981)

The literature of digital transmission on wire-pair cables generally considers the probability distributions of both pair-to-pair crosstalk loss and its power sum to be normal on a dB scale. This paper presents extensive measured data of crosstalk among connectors and wire pairs on the backplane and associated stub cable of 466-type apparatus cases of the existing T1 system. The measured probability distribution of crosstalk power sum "bends" toward more severe crosstalk levels in the lower tail region, which is important for T1 system engineering. This bend is because of the effects of a few dominant components (i.e., within-slot or within-harness crosstalk) in the power sum. The simple normal model is too optimistic by 4 dB in estimating apparatus-case-crosstalk power sum at 0.1 percentile level. This paper shows that both the Monte Carlo and the lower bound methods for power sum calculations predict this bend in close agreement with the measured data. Although apparatus-case-crosstalk power sum is worse than previously assumed, the performance of T1 system has been adequately protected by the extra margin in the previous engineering rules to cover unknowns.

I. INTRODUCTION

Crosstalk interference is a prime limitation on the transmission capacity and the performance of digital transmission systems, such as T1,[1] T1C,[2,3] SLC-40,[4] T1D,[5] and SLC-96,[6] on twisted multipair cables. An important step in the design of digital systems and their associated engineering rules is the characterization of the power sum of pair-to-pair crosstalk loss. The crosstalk power sum is the total crosstalk interference which appears on a given pair as a result of coupling from all disturbers on other pairs.

The crosstalk power sum of a T-carrier system can be decomposed into two components: one component originates from crosstalk among wire pairs in the cable, and the other component originates from crosstalk among wire pairs on the backplane of the repeater apparatus case and the associated connectors and stub cable. At each repeater location of a T1 system, an apparatus case is used to house 50 regenerators of 50 one-way T1 systems. Figure 1 shows the 25-slot arrangement of the 466-type (without lightning protection device) apparatus case. Each slot holds two T1 regenerators. Figure 2 shows a portion of the wiring arrangement between the stub cable and the repeater connectors on the backplane of a 466-type apparatus case. The crosstalk originating from the wire pairs on the apparatus case backplane, connectors, and stub cable is known as apparatus-case-crosstalk (ACXT).

In new, 800-series, plastic apparatus cases, a carefully controlled wiring layout is used to minimize the ACXT to such an extent that ACXT can be neglected in the T1 system engineering rules. However, extensive laboratory and field measurements indicate that the old vintage apparatus cases, such as 466-type, are major contributors to T1 intersystem crosstalk. In this paper, we study the ACXT data of 466-type apparatus cases in detail because this is one type of apparatus case which has been widely deployed in the existing T1 plant. The statistics of ACXT are, therefore, important in characterizing the performance of the existing T1 systems and future higher bit rate systems proposed to be used in the existing T1 environment. All the ACXT data presented in this paper were measured at 0.772 MHz, the Nyquist frequency of the T1 bit rate. As explained in Section II of Ref. 7, the extreme tail region (i.e., 0.1 to 0.025 percent) of the probability distribution of the crosstalk power sum is important in the engineering of digital transmission systems in twisted pair cables.

Figure 3 shows the distribution of the power sum of ACXT of 466-type apparatus cases obtained from extensive laboratory[8] and field measurements.[9] In the simple normal model, the power sum data on Fig. 3 would be fitted by a straight line and the predicted 0.1 percentile would be 59 dB. However, the measured power sum distribution on Fig. 3 has a noticeable "bend" towards more severe crosstalk levels in

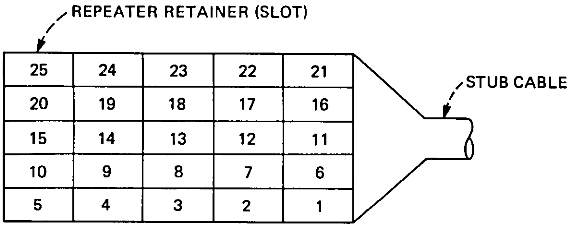


Fig. 1—The 466-type apparatus case with 25 repeater slots (i.e., retainers) for T1 repeaters.

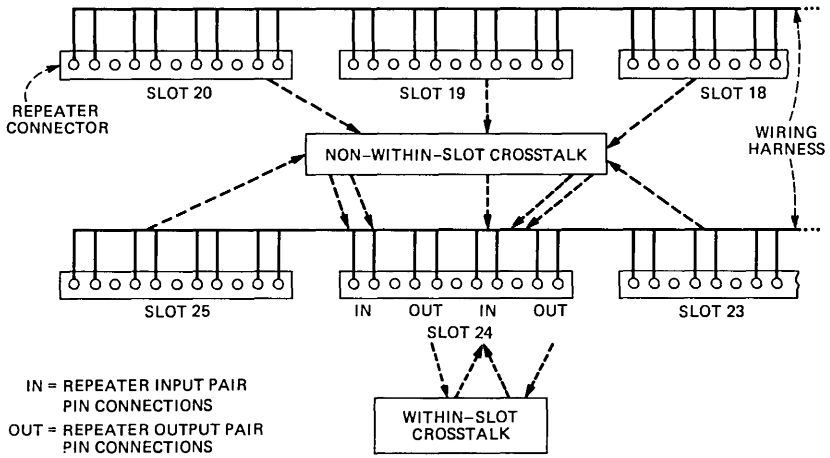

REPEATER CONNECTOR
SLOT 20
SLOT 19
SLOT 18
WIRING HARNESS
NON-WITHIN-SLOT CROSSTALK
SLOT 25
IN
OUT
IN
OUT
SLOT 24
SLOT 23
IN = REPEATER INPUT PAIR PIN CONNECTIONS
OUT = REPEATER OUTPUT PAIR PIN CONNECTIONS
WITHIN-SLOT CROSSTALK

Fig. 2—Wiring diagram on the backplane of 466-type T1 apparatus case showing within-slot crosstalk and non-within-slot crosstalk for slot 24.

the lower tail region (≤5 percent). It cannot be described precisely by any simple model such as normal or gamma. This paper shows that the bend is because of the dominant effect of the within-slot pair-to-pair crosstalk which is, on the average, 15 dB worse than the non-within-slot pair-to-pair crosstalk. It is demonstrated that both the

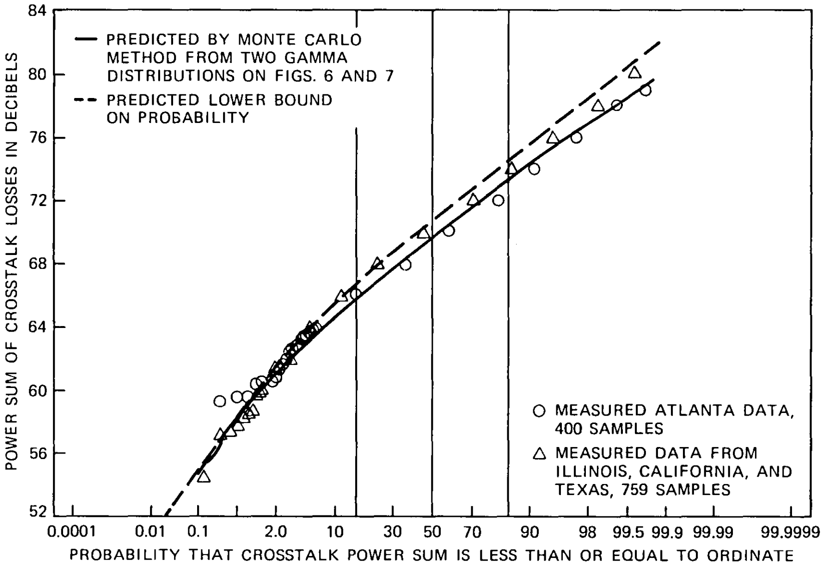


Fig. 3—Distribution of power sum of 50 pair-to-pair crosstalk losses of 466-type T1 apparatus case from laboratory measurements in Atlanta and field measurements in Illinois, California, and Texas.

Monte Carlo and the lower bound method for power sum calculations predict this bend. Both methods and the data indicate that the 0.1 percentile of the power sum distribution for 466-type cases is 55 dB which is 4 dB more pessimistic than the 59 dB predicted by the simple normal model. Therefore, the simple normal model may be too optimistic in estimating the T1 system margin. The previous engineering rules[10] for T1 system contain extra margin to cover "unknowns." The effect of ACXT on the bit-error-rate performance of T1 system has been adequately protected by the extra margin. The development of a more accurate ACXT model will reduce the unknowns and enable a greater exploitation of the system's capability by mitigating the need for large "uncertainty" margins. M. H. Meyers[11] has also investigated ACXT by a different approach.

II. ATLANTA DATA AND THE SIMPLE NORMAL AND GAMMA MODELS

The ACXT data of eight 466-type apparatus cases were measured in Bell Laboratories in Atlanta by using a computer operated transmission measurement set.[8,12,13]

The laboratory data are shown as circles on Figs. 4 and 5 for pair-to-pair crosstalk loss and the power sum, respectively. The solid line and the dashed line on Fig. 4 represent the gamma and normal approximation, respectively, to the pair-to-pair distribution. The power sum distribution is strongly controlled by the behavior of pair-to-pair distribution in the tail region of low crosstalk loss.[7] The gamma

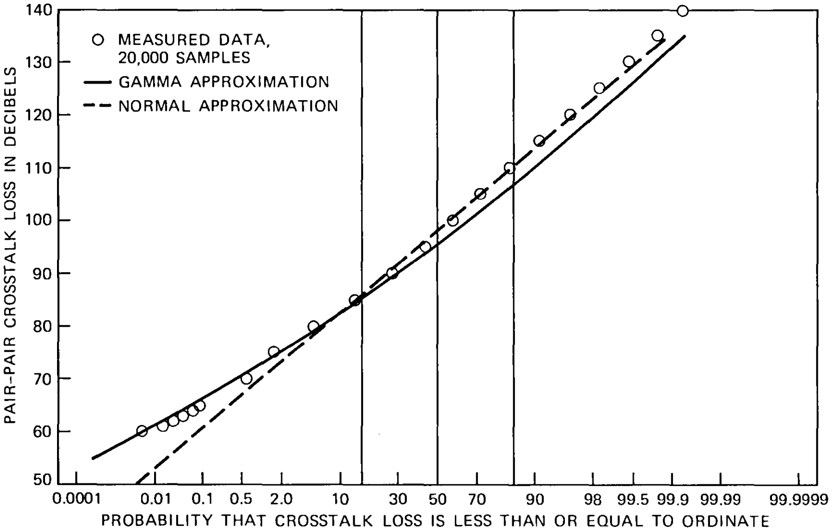


Fig. 4—Distribution of pair-to-pair crosstalk loss of 466-type T1 apparatus case. Data measured from eight apparatus cases in Bell Laboratories, Atlanta.

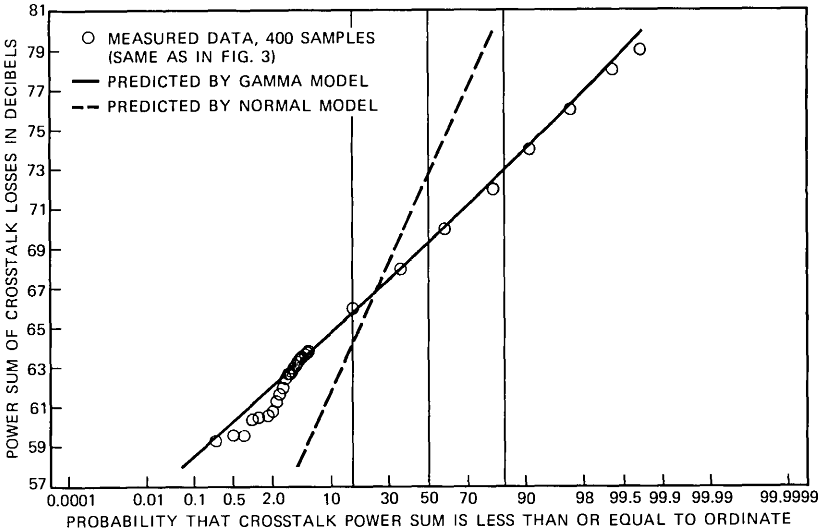


Fig. 5—Distribution of power sum of 50 pair-to-pair crosstalk losses of 466-type T1 apparatus case. Data measured from eight apparatus cases in Bell Laboratories, Atlanta.

approximation fits the pair-to-pair data very closely in the critical region of low crosstalk loss, whereas the normal approximation is too pessimistic in this important tail region.

The worst value of 60 dB on Fig. 4 does *not* imply that the pair-to-pair distribution is truncated at the 60-dB level. A finite sample measurement of a random variate (e.g., crosstalk loss) always yields a finite worst value even if the parent distribution of the variate is untruncated. The sample worst value varies randomly from one set of measurement (e.g., from one cable) to another. The probability distribution of the worst value (i.e., the extreme value) is the subject of extreme value statistics which have been studied extensively.[14,15,16] Therefore, the existence of a finite worst value from a finite sample measurement of cable crosstalk loss does not necessarily imply that the parent distribution of crosstalk loss is truncated.

Figure 5 shows that the power sum distribution predicted by the gamma model (solid line) agrees reasonably well with the measured data over a large portion of the distribution, but the discrepancy in the lower tail region is noticeable. On the other hand, the prediction by the untruncated normal model (dashed line) differs substantially from the data. The equations and the calculation procedure of the gamma model have been described in Ref. 7. The estimated statistical parameters of apparatus-case crosstalk based on the simple gamma model are listed in Table I.

Many authors have used the Wilkinson's method[17] to calculate the power sum distribution from the pair-to-pair distribution. The obvious

Table I—Statistical parameters of apparatus-case crosstalk

(Eight 466-type apparatus cases measured in Atlanta)

Gamma Model† (For all data)		
Pair-to-pair	$\bar{y}$	3.30×10^{-8}
	σ_y	2.23×10^{-8}
	ν	80.0
	β	0.833
	M_x(dB)	96.0*
	σ_x(dB)	10.7*
Power sum of 50 pair-to-pair crosstalk losses	$\bar{s}$	1.65×10^{-7}
	σ_s	1.58×10^{-7}
	μ	568.00
	α	6.58
	M_Q(dB)	69.30
	σ_Q(dB)	3.62

* These values are estimated by gamma model and are slightly different from those of normal model.

† The definitions of terms and equations related to gamma model are given in Ref. 7.

discrepancy between the measured data and the dashed line predicted by the untruncated normal model in Fig. 5 has prompted some authors to abandon the Wilkinson's method[17] entirely and to simply fit the measured power sum distribution on Figs. 3 and 5 by a normal distribution. As will be shown later, this approach is too optimistic by 4 dB at the critical 0.1 percent point. Thus, the normal model faces a dilemma of being too pessimistic (see Fig. 5), if Wilkinson's method of power sum calculation is used, and being too optimistic at the 0.1 percent point, if Wilkinson's method is by-passed (i.e., simply fit the measured power sum distribution by a normal distribution). The use of truncated normal model for pair-to-pair distribution suffers a drawback of uncertain truncation point as discussed in Ref. 7 and several dBs of error at the 0.1 percent point just mentioned.

In engineering applications, the behavior of the power sum distribution in the lower tail region is most important because the engineering objective of T1 systems is set at the 0.1 percent point for 50-section metropolitan applications. Unfortunately, Figs. 3 and 5 show that the measured data deviate substantially from the predictions by gamma and normal models in the important lower tail region. These discrepancies are predictable by both the Monte Carlo and the lower bound method as discussed in the next section.

Reference 7 and this paper indicate that an accurate prediction of crosstalk power sum distribution from pair-to-pair distribution is often difficult. One is tempted to abandon the pair-to-pair crosstalk statistics entirely and to rely solely on the measured power sum distribution. However, the studies of pair-to-pair statistics and other decompositions, such as within-slot versus non-within-slot crosstalk, and within-harness versus non-within-harness crosstalk, are necessary to provide insights for understanding and for proper modeling of non-Gaussian power sum distribution. The technique for predicting power sum distribution from pair-to-pair distribution is also necessary in characterizing some practical situations where the apparatus cases are only partially filled.

III. POWER SUM CALCULATIONS BY MONTE CARLO AND LOWER BOUND METHODS

The extensive laboratory and field measurements indicate that the distribution of the power sum of apparatus-case crosstalk has a noticeable bend towards more severe crosstalk levels in the lower tail region as shown in Fig. 3. The lower tail region of the measured data on Fig. 3 has an effective standard deviation (i.e., slope) of 6 dB on the normal probability coordinates. This slope agrees very well with the slope of the measured distribution of T1 repeater section margins in the lower tail region.[9] Thus, the laboratory measurements and field measurements consistently indicate that the distribution of the power sum of ACXT cannot be described precisely by a simple model, such as the normal or the gamma distributions.

With such understanding, we will avoid assuming any simple model for the total power sum distribution and will use more sophisticated techniques, such as the Monte Carlo method or the lower bound technique to obtain the correct power sum distribution.

Previous studies by Marlow[18] and Janos[19] indicate that the power sum distributions will have a bend if there are strong, dominant components whose mean or standard deviation differs substantially from those of the other components of the power sum. Under such circumstances, the lower tail of the power sum distribution will behave like that of the dominant components and, hence, show a bend. With this hint, we naturally look for the possible existence of dominant components in the power sum of apparatus-case crosstalk disturbers.

Figures 1 and 2 show the repeater slot and wiring arrangements of T1 apparatus case. Each T1 system in an apparatus case suffers from two within-slot disturbers* and 48 non-within-slot disturbers assuming the case is 100 percent filled. The Atlanta laboratory data show that the mean value of the within-slot crosstalk is 15 dB worse than that of the non-within-slot crosstalk. Such a large difference means that the within-slot crosstalk and the non-within-slot crosstalk must be treated separately in the power sum calculations.

The circles on Fig. 6 show the distribution of power sum of the 48 non-within-slot crosstalk disturbers measured in Atlanta Laboratory.

*Each apparatus case slot holds two T1 regenerators.

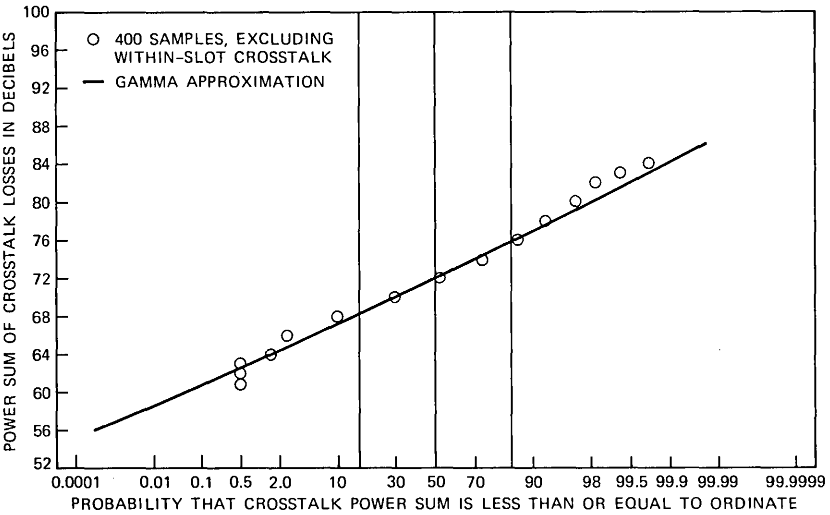


Fig. 6—Distribution of power sum of 48, pair-to-pair, non-within-slot, crosstalk losses of 466-type T1 apparatus case. Data measured from eight apparatus cases in Bell Laboratories, Atlanta. Data represents between-slot crosstalk component of data in Fig. 5.

The circles, triangles, and crosses on Fig. 7 show the distributions of power sum of the two within-slot-crosstalk disturbers measured in Atlanta, Illinois, and California, respectively. The solid lines on Figs. 6 and 7 are the corresponding gamma approximations with the param-

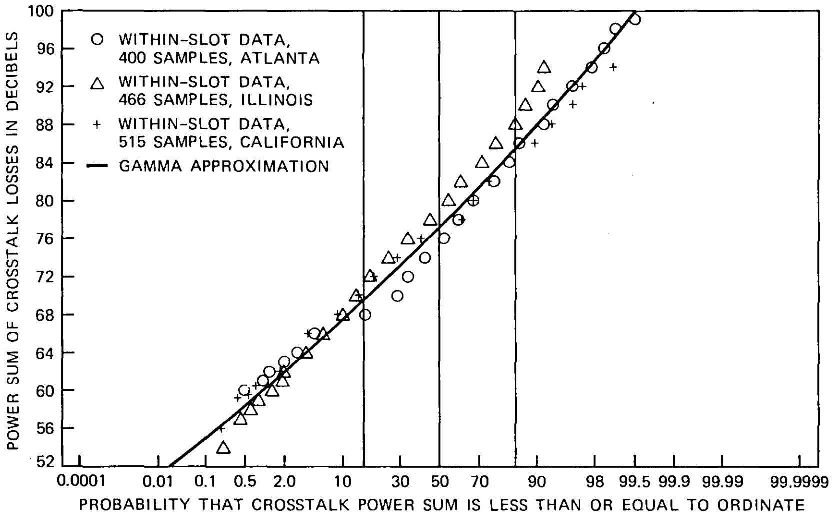


Fig. 7—Distributions of power sum of two, pair-to-pair, within-slot, crosstalk losses of 466-type T1 apparatus case from laboratory measurements in Atlanta and field measurements in Illinois and California. Data represents within-slot crosstalk component of data in Figs. 3 and 5.

Table II—Statistical parameters of non-within-slot crosstalk of apparatus case

(Eight 466-type apparatus cases measured in Atlanta)

	Gamma model for power sum of 48 non-within-slot crosstalk (Fig. 6)
$\bar{s}$	8.99×10^{-8}
σ_s	9.10×10^{-8}
μ	551.00
α	6.20
M_Q(dB)	72.20
σ_Q(dB)	3.79

eters listed in Tables II and III, respectively. The total power sum of the 50 pair-to-pair crosstalk disturbers is equal to the power sum of the two gamma distributed random variables on Figs. 6 and 7. Tables II and III show that the mean values of these two gamma variates differ by only 7 percent (i.e., 72.2 vs. 77.5 dB), whereas the standard deviations differ by 100 percent (i.e., 3.8 versus 8.0 dB). A bend on the distribution of their power sum is, therefore, expected.

The Monte Carlo method for power sum calculation has been used by many authors.[17,18] The lower bound technique is described in the Appendix. Figure 3 shows that both the Monte Carlo and the lower bound methods predict the bend in the total power sum distribution in close agreement with the measured data. The Monte Carlo result* agrees very well with the measured data over the entire range. The predicted lower bound (in probability) is practically identical to the Monte Carlo result in the lower tail region ($\leq$2 percent) and is applicable to T1 system engineering. The lower bound method has the advantages of being simple computationally and of providing some physical insights into the power sum behavior in the tail region as discussed below.

Let x denote the power sum resulting from within-slot crosstalk (Fig. 7) and let y denote the power sum due to non-within-slot crosstalk (Fig. 6). Furthermore, let z denote the power sum of x and y, the total power sum of 50 pair-to-pair crosstalk disturbers. The Appendix shows that a lower bound, $P_{LB}(z \leq b)$, of the probability distribution of z is:

$$P_{LB}(z \leq b) = P(x \leq b) + P(y \leq b) - P(x \leq b) \cdot P(y \leq b), \qquad (1)$$

where b represents the crosstalk level at which the probabilities are of interest. The data in Figs. 6 and 7 show that

$$P(y \leq b) \ll P(x \leq b) \quad \text{for} \quad b \leq 62 \text{ dB}, \qquad (2)$$

*A sample size of 5000 is used in obtaining the Monte Carlo result (the solid line) in Fig. 3.

Table III—Statistical parameters of within-slot crosstalk of apparatus case
(466-Type apparatus cases measured in Atlanta, Illinois, and California)

Gamma model for power sum of two within-slot crosstalk (Fig. 7).	
$\bar{s}$	8.45×10^{-8}
σ_s	2.75×10^{-7}
μ	101.00
α	1.258
M_Q(dB)	77.50
σ_Q(dB)	8.00*

* Since a gamma distribution is not a straight line on a normal probability coordinates, the slope of the gamma distribution in the lower tail region is not the same as that (i.e., 8 dB of σ) in the middle range.

which implies that:

$$P_{LB}(z \le b) \simeq P(x \le b) \quad \text{for} \quad b \le 62 \text{ dB}. \tag{3}$$

The data in Figs. 3 and 7 indeed support this simple approximation. Therefore, the lower bound method demonstrates through eq. (3) that the total power sum distribution behaves like that of the dominant component x in the lower tail region (i.e., for those situations where ACXT is worse than 62 dB). Notice that the dominant component x represents the power sum of the two within-slot-crosstalk losses.

VI. CONCLUSION

The extensive data on T1 apparatus-case crosstalk for 466-type cases from laboratory measurement in Atlanta and field measurements in Illinois, California, and Texas consistently indicate that the power sum distribution has a bend towards more severe crosstalk levels in the lower tail region. It is shown that this bend is because of the dominant effect of within-slot crosstalk. Both the Monte Carlo and the lower bound methods of power sum calculations predict this bend if the power sum contains a dominant component which differs substantially from other components.

The 0.1 percentile of the distribution of power sum of ACXT is about 55 dB for 466-type apparatus case. This is about 4 dB worse than that predicted by the conventional normal model which ignores the bend in the tail region.

VII. ACKNOWLEDGMENT

I am grateful to L. T. Nguyen for providing the Monte Carlo analysis. The laboratory data were measured by J. Kreutzberg and his group at

Bell Laboratories in Atlanta. The extensive field measurements of T-carrier characteristics, including ACXT, were conducted by D. G. Bean, R. L. Brown, K. J. Burns, J. R. Davis, G. F. Erbrecht, J. P. Fitzsimmons, L. E. Forman, J. H. Gentry, T. C. Kaup, D. G. Leeper, W. J. Maybach, M. H. Meyers, A. K. Reilly, P. E. Scheffler, and R. J. Schweizer.

APPENDIX

A Lower Bound of Power Sum Distribution

This appendix describes a technique to obtain a lower bound of the probability distribution of crosstalk power sum. In the low probability tail region, this lower bound is fairly tight and provides a simple approximation to the power sum distribution. This approach is inspired by the work of Marlow and Farley.[18,20]

Let $P(x \le b)$ and $P(y \le b)$ be the cumulative distributions of two, positive, independent, random variables x and y, respectively, and let

$$z = -10 \log_{10}[10^{\frac{-x}{10}} + 10^{\frac{-y}{10}}] \tag{4}$$

be the power sum of x and y. This definition implies that

$$z \le \min(x, y), \tag{5}$$

and

$$P(z \ge b) \le P(\min(x, y) \ge b) \tag{6}$$

$$= P(x \ge b, y \ge b), \tag{7}$$

where $\min(x, y)$ denotes the minimum of x and y, and $P(x \ge b, y \ge b)$ denotes the probability that both x and y exceed b. The independence of x and y implies that

$$P(x \ge b, y \ge b) = P(x \ge b) \cdot P(y \ge b). \tag{8}$$

By definition:

$$\begin{aligned} P(z < b) &\equiv 1 - P(z \ge b) \\ P(x < b) &\equiv 1 - P(x \ge b) \\ P(y < b) &= 1 - P(y \ge b) \quad . \end{aligned} \tag{9}$$

Combining eqs. (7), (8), and (9) yields

$$\begin{aligned} P(z < b) &\ge 1 - P(x \ge b) \cdot P(y \ge b) \\ &= 1 - [1 - P(x < b)] \cdot [1 - P(y < b)] \\ &= P(x < b) + P(y < b) - P(x < b) \cdot P(y < b). \end{aligned} \tag{10}$$

Therefore, the right-hand side of eq. (10) represents a lower bound (in probability) for the distribution of the power sum z. Notice that this lower bound can easily be calculated when the distributions $P(x \leq b)$ and $P(y \leq b)$ of the components x and y are known.

In the lower tail region where both $P(x \leq b)$ and $P(y \leq b)$ are small, the probability of both x and y being less than b simultaneously is extremely small. Therefore, the inequality eqs. (5) and (10) asymtotically approach equalities in the lower tail region (i.e., when b is small). This means that in the low probability tail region, the lower bound eq. (10) is fairly tight and provides a simple but accurate approximation to the power sum distribution.

REFERENCES

1. K. F. Fultz and D. B. Penick, "The T1 Carrier System," B.S.T.J., *44,* No. 7 (September 1965), pp. 1405–51.
2. "TIC-A New Digital System For Paired Cable Application," IEEE International Conference on Communications, Conference Record, Session 39 (six papers), San Francisco, California, June 16–18, 1975, pp. 39-1–28.
3. R. E. Maurer, J. A. Lombardi, and J. B. Singleton, "TIC-A New Digital System for Exchange Area Applications," Proc. IEEE Nat. Telecommun. Conf., San Diego, California, December, 1974, pp. 636–40.
4. S. J. Brolin and G. E. Harrington, "The SLC-40 Digital Carrier Subscriber System," 1975 IEEE Intercon. Conference Record, Session 8—Loop Electronics—The New Frontier, International Convention and Exposition, New York, April 8-10, 1975.
5. "T1D Digital Transmission System," IEEE National Telecommunications Conference, Session 39 (six papers), Houston, Texas, November 30-December 4, 1980, pp. 39.1.1–6.3.
6. S. Brolin et al., "Inside the New Digital Subscriber Loop System," Bell Laboratories Record, *58,* No. 4 (April 1980), pp. 110–16.
7. S. H. Lin, "Statistical Behavior of Multipair Crosstalk," B.S.T.J., *59,* No. 6 (July-August 1980), pp. 955–74.
8. J. M. Pace, private communication.
9. D. G. Leeper and P. E. Scheffler, private communication.
10. H. Cravis and T. V. Crator, "Engineering of T1 Carrier System Repeatered Lines," B.S.T.J., *42,* No. 2 (March, 1963), pp. 431-86.
11. M. H. Meyers, unpublished work.
12. R. E. Anderson, "Computer Controlled Cable Measurements," Proc. 21st International Wire and Cable Symposium, Cherry Hill, New Jersey, December 1972, pp. 188–192, Sponsored by the U. S. Army Communications Research and Development Command, Fort Monmouth, New Jersey.
13. W. J. Geldart, G. D. Haynie, and R. G. Schleich, "A 50 Hz-250 MHz Computer-Operated Transmission Measuring Set," B.S.T.J., *48,* No. 5 (May–June 1969), pp. 1339–81.
14. E. J. Gumbel, *Statistics of Extremes,* New York: Columbia University Press, 1958.
15. E. J. Gumbel, *Statistical Theory of Extreme Values and Some Practical Applications,* Applied Mathematics Series No. 33, Washington, D.C.: National Bureau of Standards, 1954.
16. H. Cramer, *Mathematical Methods of Statistics,* Princeton, New Jersey: Princeton University Press, 1974, Section 28.6.
17. I. Näsell, "Some Properties of Power Sums of Truncated Normal Random Variables," B.S.T.J. *XLVI,* Part II, No. 9 (November 1967), pp. 2091–110.
18. N. A. Marlow, private communication.
19. W. A. Janos, "Tail of the Distribution of Sums of Lognormal Variates," IEEE Trans. Inf. Theory, *IT-16,* No. 3 (May 1970), pp. 299–302.
20. J. E. Farley, private communication.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 60, No. 7, September 1981
Printed in U.S.A.

Current-Carrying Capacity of Fine-Line Printed Conductors

By A. J. RAINAL

(Manuscript received February 23, 1981)

This paper presents simple equations, along with experimentally determined parameters, to calculate the transient temperature rise of current-carrying, fine-line (~7 mils) conductors on various styles of circuit packs. The styles of circuit packs include wire wrap, double-sided, metal, bonded, and various multilayer boards. All styles of circuit packs in the BELLPAC™ system family are included. The maximum steady-state temperature rise of a nicked or constricted current-carrying conductor is also treated. The calculated transient and steady-state temperature rises agree with experimental results.

I. INTRODUCTION

Printed wiring technology presently provides the physical designer with fine-line copper conductors to interconnect integrated circuits and other components at the circuit-pack (CP) level. These fine-line conductors now have a nominal width of 7 mil, and a nominal thickness of 1.4 mil. For such relatively small conductor sizes, (~AWG39), the current-carrying capacity of the conductors becomes an important matter of concern. Can such fine-line conductors carry the required current to operate the various components on a CP without causing an excessive temperature rise? What is the temperature rise during normal current flow? If a fault occurs and a current of 10 A flows for 100 ms, will the CP be damaged? Questions of this nature are becoming very important as printed wiring technology provides finer conductors for the electrical interconnections.

Also, as large-scale integrated (LSI) circuits are introduced, the assembled CPs are becoming more expensive. Therefore, there is more incentive to protect the CP from possible damage as a result of an over current.

Some early work by W. Aung and A. J. Colucci[1] has shown that a 7.5-A current flowing through a fine-line printed conductor inside a multilayer board (MLB) can cause the MLB to break into flames in about 5 s. Also, W. T. Smith[2] reported that a current flow through a fine-line printed conductor (2.8-mils thick) should be limited to about 5 A if temperature rises are to be limited to 120°C.

Clearly, to avoid possible disaster or damage, the physical designer of electronic equipment must be able to estimate the transient temperature rise of fine-line, current-carrying conductors on all styles of circuit packs and backplanes.

During the past few years, the Bell System has introduced a modular packaging system (*BELLPAC** system[3]) for packaging electronic equipment. This system makes use of a number of CP styles that have common features suitable for computer-aided design. Fine-line printed conductors are available for use on any of the CP styles. The *BELLPAC* system project has provided us with the opportunity to study the current-carrying capacity of fine-line conductors on a variety of CP styles.

The purpose of this paper is to present some useful results concerning the current carrying capacity of fine-line conductors on various styles of CPs. The results include the transient temperature rise of a current-carrying printed conductor, and the maximum steady-state temperature rise of a conductor which may be nicked or constricted.

For the special case of a double-sided epoxy printed wiring board (flex or rigid), some results concerning the steady-state temperature rise of fine-line printed conductors have been reported in Refs. 4 and 5. Although the methods reported in these two references differ, the results agree well with one another.

A listing of the CP styles of interest in this paper, along with a short description of each is presented in Table I. Figure 1 shows the corresponding physical layups of the CP styles. These physical layups include all of the CP styles presently in the *BELLPAC* system. We shall see that the results in this paper can be applied to estimate the current-carrying capacity of fine-line conductors on any layer of any of the CP styles shown in Fig. 1.

II. BASIC EQUATIONS

2.1 General results

For a general current-carrying conductor, the conservation of heat energy requires that the average temperature rise satisfy the following differential equation:

$$I^2R_1[1 + \alpha_1\overline{\Delta T}]\mathrm{dt} = \mathrm{C}\mathrm{d}\overline{\Delta T} + \frac{\overline{\Delta T}}{R_T}\,\mathrm{dt}, \tag{1}$$

* Trademark of Western Electric.

Table I—Description of the circuit-pack styles

Circuit-Pack Style	Description
Wire wrap	Wire-wrap board for breadboarding
Extender board	6-Layer MLB, 2-pad layers, 2 signal layers, power (P) and ground (G) on inside, dedicated ground conductor between every pair of signal conductors
Double-sided (epoxy)	Double-sided, epoxy PWB
Double-sided (metal)	Double-sided, metal core, PWB
Bonded board (LAMPAC)*	Flex bonded to epoxy-coated steel
4L MLB (EXT P/G)	4-Layer MLB, 2 signal layers, P and G on outside
6L MLB (EXT P/G)	6-Layer MLB, 4 signal layers, P and G on outside
6L MLB (INT P/G)	6-Layer MLB, 2 pad layers, 2 signal layers, P and G on inside
6L MLB (INT P/G, Surface Routing)	6-Layer MLB, 4 signal layers, P and G on inside
8L MLB (INT P/G)	8-Layer MLB, 2 pad layers, 4 signal layers, P and G on inside

* This particular bonded board is also known as LAMPAC.

where

I = current flow through the conductor (amperes)
R_1 = resistance of the conductor at ambient temperature (ohms)
α_1 = ambient temperature coefficient (per degree C)
= $[T_1 + 234.45]^{-1}$ (a good approximation in the case of copper)
T_1 = ambient temperature (°C)
t = time duration of the current flow (seconds)
C = thermal capacity of the material heated (J/°C)
R_T = thermal resistance of the conductor on the CP style of interest (°C/W)
$\overline{\Delta T}$ = average temperature rise (°C) along the length of the conductor.

The left-hand side of eq. (1) represents the energy dissipated, and the two terms on the right-hand side represent the stored and radiated energy, respectively. More general forms of eq. (1) are discussed in Ref. 6.

Laboratory measurements show that the thermal capacity, C, is time dependent. The physical reason for this dependence is that more and more of the material is heated as time goes on. Initially, only the conductor and a small portion of the substrate and covercoat are heated.

However, if the time axis is partitioned into appropriate time intervals, it turns out that the thermal capacity, C, is approximately constant over each of the time intervals. The solution to eq. (1) in the case of three such time intervals is given by:

$$\overline{\Delta T} = \overline{\Delta T}_{ss}\,[1 - \exp\{-(R_T C_1)^{-1}(1 - I^2 R_1 R_T \alpha_1)t\}] \qquad 0 \le t \le t_1 \qquad (2)$$

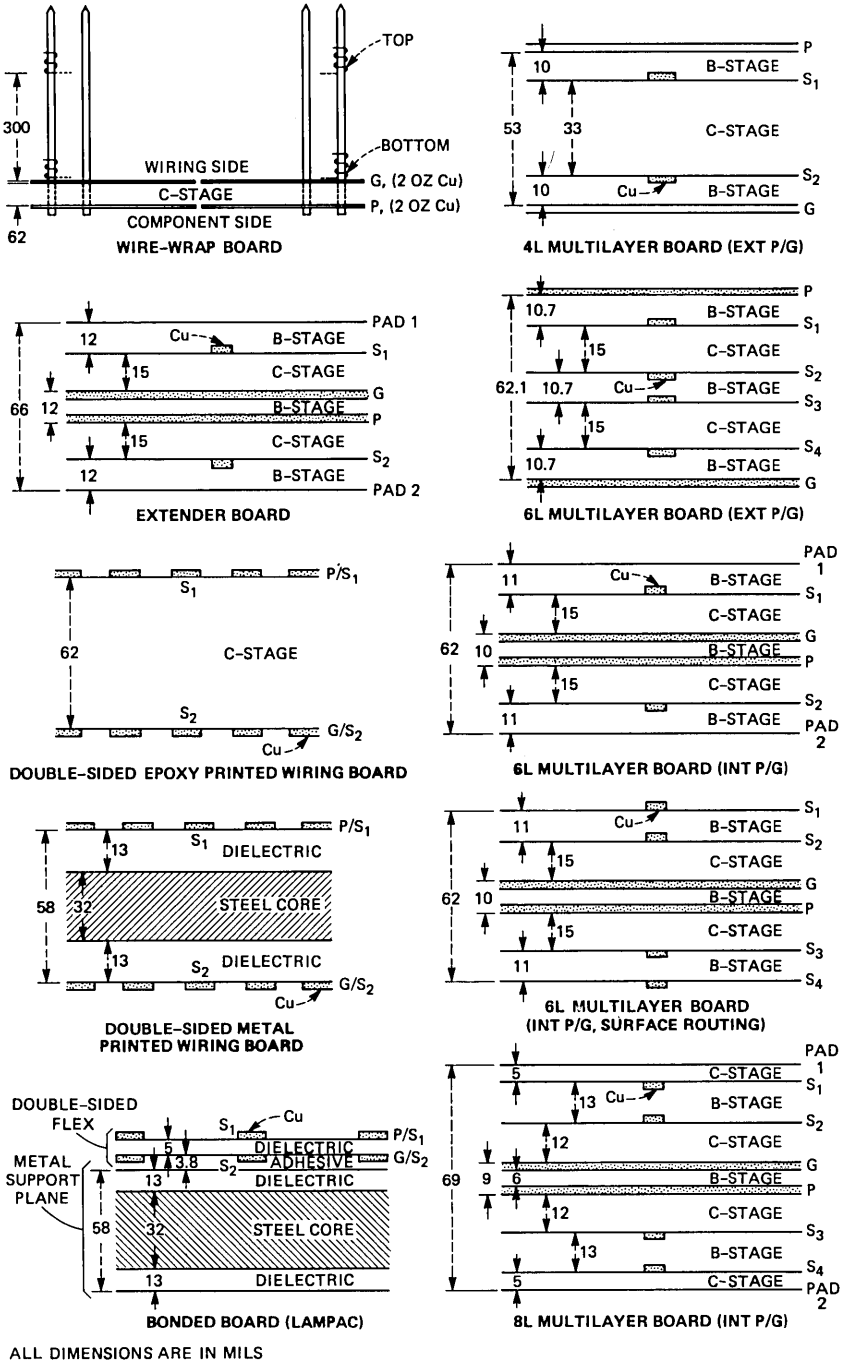


Fig. 1—Lay-ups of the various circuit-pack styles.

$$\overline{\Delta T} = \overline{\Delta T}_{ss}\left[1 - \exp\left\{-(R_T C_1)^{-1}(1 - I^2 R_1 R_T \alpha_1)\right.\right.$$
$$\left.\left.\cdot\left(1 - \frac{C_1}{C_2} + \frac{C_1}{C_2}\frac{t}{t_1}\right) t_1\right\}\right] \qquad t_1 \le t \le t_2 \tag{3}$$

$$\overline{\Delta T} = \overline{\Delta T}_{ss}\left[1 - \exp\left\{-(R_T C_1)^{-1}(1 - I^2 R_1 R_T \alpha_1)\right.\right.$$
$$\left.\left.\cdot\left(1 - \frac{C_1}{C_2} + \frac{C_1 t_2}{C_2 t_1} - \frac{C_1 t_2}{C_3 t_1} + \frac{C_1}{C_3}\frac{t}{t_1}\right) t_1\right\}\right] \qquad t_2 \le t \le \infty, \tag{4}$$

where

$\overline{\Delta T}_{ss} = \dfrac{I^2 R_1 R_T}{(1 - I^2 R_1 R_T \alpha_1)}$ = steady-state average temperature rise,

C_1 = thermal capacity during the time interval $[0, t_1]$,

C_2 = thermal capacity during the time interval $[t_1, t_2]$,

C_3 = thermal capacity during the time interval $[t_2, \infty]$.

In general, one can partition the time axis into n contiguous time intervals and obtain a set of n equations. For our purposes, $n = 3$ proved to be sufficient.

The current-carrying capacity of a conductor is limited by the permissible temperature rise of the conductor above the ambient temperature. Therefore, once the pertinent parameters are known, the above equations can be used to calculate the current-carrying capacity of a particular conductor.

In Section III, we describe the experimental method used to measure all of the pertinent parameters.

2.2 Some special cases

2.2.1 Runaway or critical current

The functional form of $\overline{\Delta T}_{ss}$ shows that a runaway or critical current, I_c, exists for which $\overline{\Delta T}_{ss} \to \infty$. That is, as $I \to I_c$, $\overline{\Delta T}_{ss} \to \infty$, and the current-carrying conductor never reaches steady-state temperature. The value of I_c is given by:

$$I_C = \frac{1}{\sqrt{R_1 R_T \alpha_1}}. \tag{5}$$

The phenomenon of runaway and the value of runaway current is consistent with our experience in the laboratory. We found that as we approached the critical value, I_c, the temperature of the conductor rises rapidly beyond the tolerable limits of the substrate and permanent damage results.

2.2.2 Small t — initial temperature rise

From eq. (2), we find that as $t \to 0$, we have

$$\overline{\Delta T} = \frac{I^2 R_1 t}{C_1}. \tag{6}$$

Notice that this result is independent of the thermal resistance R_T. That is, the initial heating process is adiabatic.

2.2.3 *Large t — steady state temperature rise*

From eq. (4), with $I < I_c$, we see that as $t \rightarrow \infty$ we have

$$\overline{\Delta T} = \overline{\Delta T}_{ss} = \frac{I^2 R_1 R_T}{(1 - I^2 R_1 R_T \alpha_1)}. \tag{7}$$

Equation (7) shows that the steady-state temperature rise depends on the product of the electrical resistance (a property of the conductor) and the thermal resistance (a property of the environment of the conductor). For a given current I, and ambient T_1, one can only reduce $\overline{\Delta T}_{ss}$ by reducing the product $R_1 R_T$.

III. EXPERIMENTAL DETERMINATION OF THE PARAMETERS

In order to carry out this study, appropriate test boards were designed for each CP style shown in Fig. 1. Except for the double-sided metal board, all test boards were fabricated at the Western Electric printed-circuit manufacturing plant at Richmond, Virginia. The double-sided metal board was manufactured at the Western Electric plant in Kearny, New Jersey.

In all cases, the physical dimensions of the printed conductors had the nominal values of length L = 12 in., width W = 7 mil, and copper thickness t_0 = 0.5 to 3 mil.

The experimental method used to determine the pertinent parameters is based on measuring, indirectly, the average temperature rise, $\overline{\Delta T}$, along the conductor as a function of time when a step function of current is applied. The measurement is based on the well-known resistance thermometer formula:[6,7]

$$\frac{V}{I} = R = R_1[1 + \alpha_1 \overline{\Delta T}], \tag{8}$$

where

V = voltage across the conductor,
I = magnitude of the step function of current,
R = measured resistance of the conductor.

The procedure used to measure $\overline{\Delta T}$ is as follows: The ambient temperature T_1 is recorded and R_1 is measured by means of an ac Kelvin bridge. This measurement involves a small current (100 mA or less) which causes a negligible temperature rise. Then a step function of I amperes is directed through the conductor of interest, and the resulting voltage drop, V, across the conductor is recorded as a function of time. Equation (8) is then used to deduce the corresponding $\overline{\Delta T}$ as a function of time.

To help ease the data gathering, a Kaye Instruments digistrip transmitter was used to format the data for print out and magnetic tape storage on a Texas Instruments model 733 data terminal. Subsequently, the data was transmitted (via an acoustical coupler) over the telephone line to the computation center for storage. At this point, special computer programs were used to edit the stored data and to produce the computer plots.

The procedure used to determine the constants appearing in eqs. (2), (3), and (4) is as follows: The average temperature rise is first computed from eq. (8) by using the steady state voltage value corresponding to a current flow of I amperes for a sufficiently long time (usually about 10 min). This is repeated for a number of different values of current. Then, the best value (minimum mean-square-error sense) of the product R_1R_T is determined from the measured data and eq. (7) which can be rewritten as

$$\frac{\overline{\Delta T}_{ss}}{1+\alpha_1\overline{\Delta T}_{ss}} = R_1R_TI^2. \tag{9}$$

If the left-hand side of eq. (9) is plotted as a function of I^2, the slope, R_1R_T, of the best fitting line is the quantity of interest. Since the value of R_1 is known, R_T can be determined.

According to eqs. (2), (3), and (4), a plot of the measured values of

$$Y \equiv -\ln\left[1 - \frac{\overline{\Delta T}}{\overline{\Delta T}_{ss}}\right]$$

versus time, t, yields points which tend to fit a series of approximately broken lines of positive slope. From this plot, values of t_1 and t_2 can be selected as the break points of these broken lines. In the region $0 \leq t \leq t_1$, the best value of C_1 (minimum mean-square-error sense) is determined by equating the slope of the best fitting line to the slope of the negative of the exponent in eq. (2). In a similar manner, C_2 is determined by using the measured data and the slope of the negative of the exponent in eq. (3). Finally, C_3 is determined by using the measured data and the slope of the negative of the exponent in eq. (4). At this point, all of the parameters needed in eqs. (2), (3), and (4) are known, and these equations can be used to calculate the average temperature along the conductor as a function of applied current, time, conductor resistance, and ambient temperature.

In this manner, the pertinent parameters were determined for fine-line conductors on all signal layers of all CP styles shown in Fig. 1. The resulting parameters are presented in Table II. The values of $t_1 = 0.55$ s and $t_2 = 3.55$ s were found to apply to all of the CP styles. The results were scaled to a conductor length of 12 in. and width of 7 mil by using the scaling laws $C_i \sim L$, $R_T \sim 1/L$, and $H \sim 1/W$.

The values of H listed in Fig. 2 and Table II are steady-state parameters and will be discussed in more detail in Section V.

IV. EXPERIMENTAL VERIFICATION

Figure 2 presents the experimental values of $Y(t)$ for the case of a double-sided epoxy printed wiring board (PWB) with covercoat. The parameters t_1, t_2, C_1, C_2, C_3, and R_T were determined by the methods described in Section III. The final C_i were determined by averaging the results over three different values of current.

For the double-sided epoxy CP, Fig. 3 compares the experimental values of the transient temperature rise, $\overline{\Delta T}(t)$, with the corresponding theoretical values. The theoretical values were determined by using

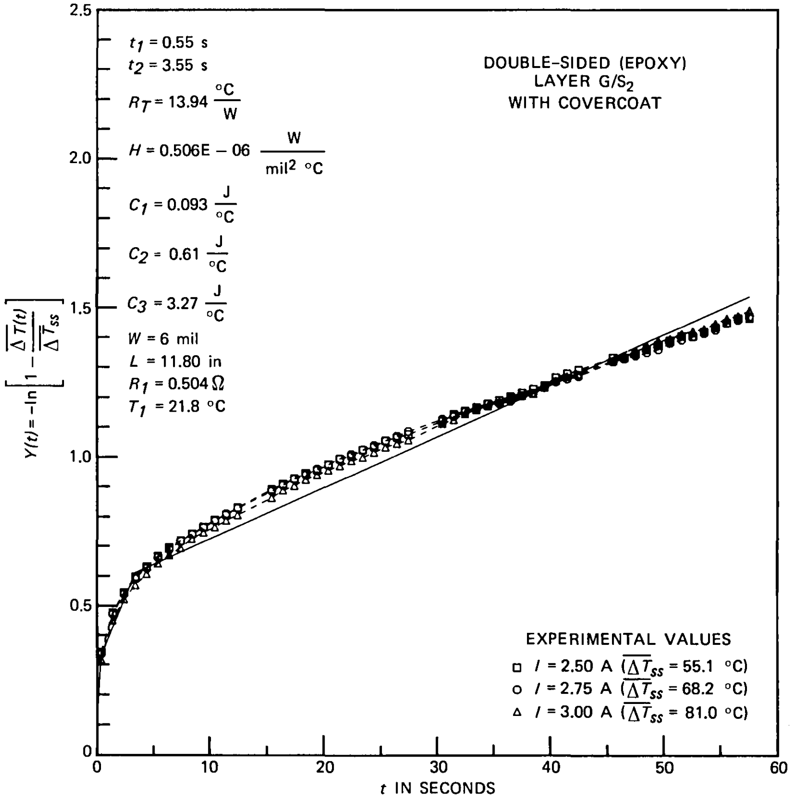


Fig. 2—Experimental values of $Y(t)$. Slopes of the broken lines determine the average thermal capacities, C_i, in the three time intervals.

Table II—Measured parameters for various circuit-pack styles (Conductor length = 1 ft, conductor width = 7 mil, and wire-wrap diameter = 10 mil)

Circuit-Pack Style*	$C_1 \frac{\text{J}}{^\circ\text{C}}$	$C_2 \frac{\text{J}}{^\circ\text{C}}$	$C_3 \frac{\text{J}}{^\circ\text{C}}$	$R_T \frac{^\circ\text{C}}{W}$	$H \frac{\text{Watts}}{\text{mil}^2\ ^\circ\text{C}}$
Wire wrap					
(Milene)†	0.118	0.170	0.265	30.5	$0.869\ 10^{-7}$
(Teflon)‡	0.150	0.259	0.406	30.2	$0.877\ 10^{-7}$
Extender board	0.121	0.697	8.71	6.32	$0.942\ 10^{-6}$
Double-sided (epoxy)	0.095	0.623	3.32	13.71	$0.434\ 10^{-6}$
Double-sided (metal)	0.052	0.305	11.22	9.00	$0.661\ 10^{-6}$
Bonded board $P/S1$	0.083	0.311	7.26	8.27	$0.720\ 10^{-6}$
(LAMPAC) $G/S1$	0.090	0.415	9.07	7.35	$0.810\ 10^{-6}$
4L MLB (EXT P/G)	0.124	1.06	3.65	11.49	$0.518\ 10^{-6}$
6L MLB (EXT P/G)					
S_1, S_4	0.124	1.06	3.65	11.49	$0.518\ 10^{-6}$
S_2, S_3	0.179	1.38	4.75	9.55	$0.623\ 10^{-6}$
6L MLB (INT P/G)					
S_1, S_2	0.121	0.697	8.71	6.32	$0.942\ 10^{-6}$
6L MLB (INT P/G, Surface Routing)					
S_1, S_4	0.076	0.443	5.59	10.47	$0.569\ 10^{-6}$
S_2, S_3	0.112	0.887	7.61	7.33	$0.812\ 10^{-6}$
8L MLB (INT P/G)					
S_1, S_4	0.076	0.438	5.01	9.70	$0.614\ 10^{-6}$
S_2, S_3	0.108	0.888	6.87	7.38	$0.806\ 10^{-6}$

* The circuit-pack styles having surface conductors were covercoated.
† Trademark of W. L. Gore & Associates, Inc.
‡ Trademark of E. I. DuPont de Nemours, Inc.

the constants t_1, t_2, C_1, C_2, C_3, R_T in eqs. (2), (3), and (4). Figure 3, shows that this method of estimating the transient temperature rise agrees with experimental results.

Similar plots have verified that this method of estimating the transient temperature rise also agrees with experimental results for all other cases of interest in this paper.

V. SOME APPLICATIONS

5.1 *Steady state temperature rises*

The steady-state temperature rises of fine-line printed conductors, or wire-wrap conductors, can be readily computed from eq. (7). The required values of thermal resistance, R_T, are listed in Table II for a conductor length of 1 ft, printed conductor width of 7 mil, and wire-wrap diameter of 10 mil. Also, the required values of the electrical resistance, R_1, can be computed from:

$$R_1 = \frac{\rho L}{t_0 W} \text{ (printed conductor),} \tag{10}$$

or

$$R_1 = \frac{4\rho L}{\pi D^2} \text{ (wire-wrap conductor),} \tag{11}$$

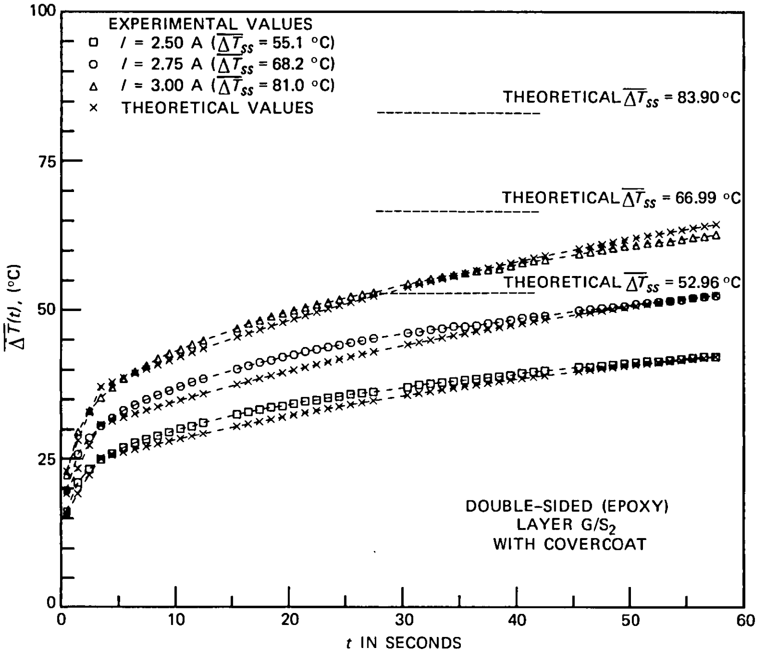


Fig. 3—Experimental values of the transient temperature rise, $\overline{\Delta T}(t)$, are compared with theoretical values.

where

$$
\begin{aligned}
\rho &= (0.67878)10^{-3}[1 + 0.00393(T_1 - 20)] \text{ ohm-mil} \\
T_1 &= \text{ambient temperature, } ^\circ\text{C} \\
L &= \text{length of conductor} = 1 \text{ ft (12,000 mil)} \\
t_0 &= \text{thickness of conductor, mil} \\
W &= \text{width of conductor} = 7 \text{ mil} \\
D &= \text{diameter of wire-wrap conductor} = 10 \text{ mil.}
\end{aligned}
$$

As an example, consider a double-sided (epoxy) CP style. From Table II, we see that $R_T = 13.71°\text{C}/W$. From eq. (10), for L = 12,000 mils, t_0 = 1.4 mils (1 oz cu), W = 7 mil and T_1 = 20°C, we find that R_1 = 0.8312Ω. For a current flow of I = 2.5 A, eq. (7) then yields $\overline{\Delta T}_{ss}$ = 98.9°C.

For conductor lengths other than L = 1 ft, one can use the scaling law $R_T \sim 1/L$ to scale the values of R_T listed in Table II. Also, R_T is essentially independent of W as is shown by eq. (16) of Ref. 4.

Let us now compute the maximum steady-state temperature rise, max ΔT_c, when the same current-carrying conductor is nicked or constricted. The maximum temperature rise of such conductors was treated in Ref. 4. It was shown that the key parameter was the value of H, the coefficient of surface heat transfer. The values of H for the CP styles of interest in this paper are listed in Table II.

As an example of the effect of a nicked or constricted conductor on steady-state temperature rise, Fig. 4 compares the computed results for a fine-line current-carrying conductor on a double-sided (epoxy) board and a 6-layer MLB (INT P/G, surface routing). At a current of $I = 2.5$ A, the nicked conductor on the double-sided board rises in temperature to about 119°C; whereas, the nicked conductor on the inside signal layer (S_2) of the 6-layer MLB rises in temperature to only 56°C.

Tabulated results for the special case $H = (0.52)10^{-6}$ (Watts/mil^2°C) were presented in Ref. 4. Since many of the values of H tabulated in Table II are close to this value, the earlier results can also be applied to many of the CP styles of interest in this paper. Also the simple relationship given as eq. (13) of Ref. 4 yields results which agree well with those presented in Fig. 4.

Table II gives the values of H when W = 7 mil. For conductor widths other than W = 7 mil, one can use the scaling law $H \sim 1/W$.

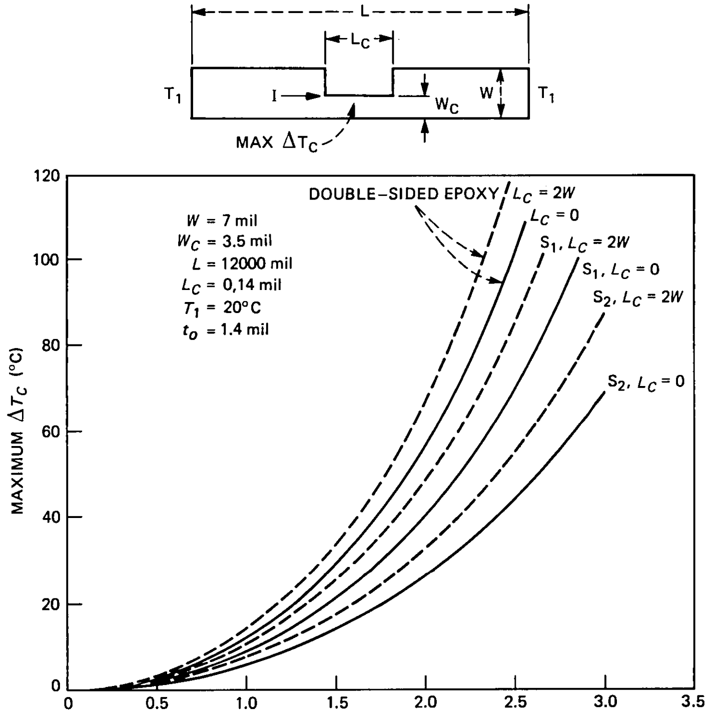


Fig. 4—The effect of a nicked or constricted conductor on the maximum steady state temperature rise.

5.2 Transient temperature rises

The transient temperature rises of fine-line printed conductors, or wire-wrap conductors can be computed from eqs. (2), (3), and (4). The required values of C_i and R_T are listed in Table II, also the values of t_1, and t_2 are given by $t_1 = 0.55$ s, and $t_2 = 3.55$ s, as was discussed in Section III.

An example of the transient temperature rise of a fine-line printed conductor on a double-sided CP is presented in Fig. 5. For this case, the appropriate parameters are listed in Table II as:

$$C_1 = 0.095 \frac{\text{J}}{^\circ\text{C}},\ C_2 = 0.623 \frac{\text{J}}{^\circ\text{C}},\ C_3 = 3.32 \frac{\text{J}}{^\circ\text{C}},\ \text{and}\ R_T = 13.71 \frac{^\circ\text{C}}{\text{Watt}}.$$

5.3 Temperature rises resulting from fault currents

A signal conductor on a CP normally carries a maximum current of about 0.1 A. Figure 4 shows that the temperature rise of such a current-

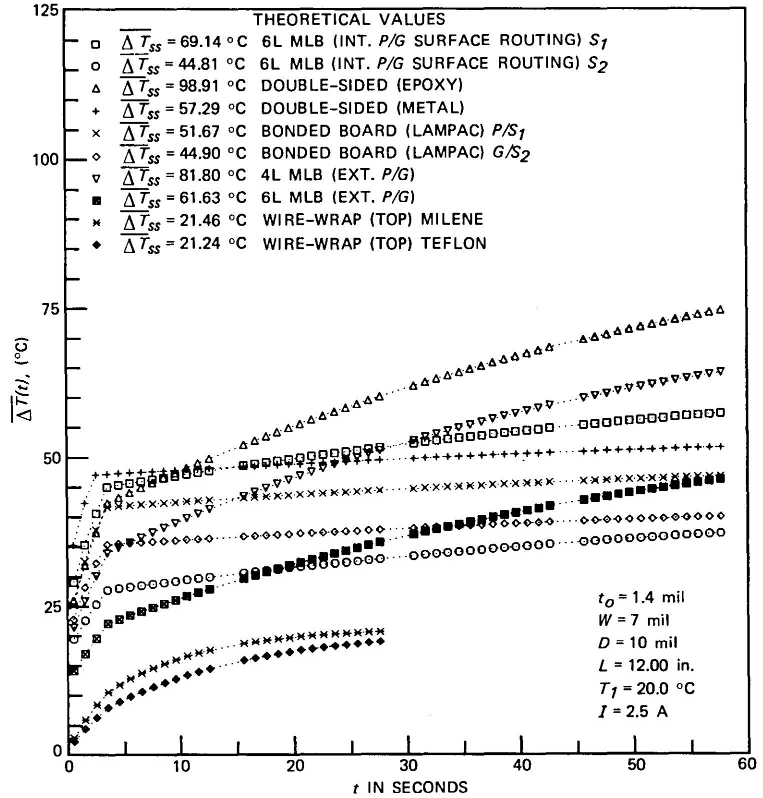


Fig. 5—Theoretical values of the transient temperature rise, $\overline{\Delta T}(t)$, are compared for different styles of circuit packs.

carrying conductor is negligible. However, a component failure or other malfunction can result in a current flow of many more amperes through the fine-line signal conductor. In this case, a circuit fuse or other over-current protection device can take many milliseconds to interrupt the flow of current. In such situations, the results in this paper can be applied to calculate the temperature rises on any layer of any of the CP styles shown in Fig. 1.

For example, consider the case of a signal conductor having the physical dimensions of $L = 1$ ft, $W = 7$ mil, and $t_0 = 1.4$ mil on a double-sided (epoxy) CP. Let us also assume that the ambient temperature is 50°C. Suppose the fault current is 10 A and the fuse or over-current protection device interrupts the fault current in 100 ms. What is the resulting temperature of the current-carrying conductor immediately before the current flow is interrupted?

From eq. (10), the electrical resistance of the conductor is $R_1 = 0.929\Omega$. From eq. (6) we find that

$$\overline{\Delta T} = \frac{I^2 R_1 t}{C_1} = \frac{(10\,A)^2(0.929\Omega)(0.1s)}{\left(0.095\,\frac{\text{J}}{°\text{C}}\right)} = 97.8°\text{C}. \tag{12}$$

The value of C_1 was taken from Table II. Thus, the temperature of the signal conductor is about 50°C + 98°C = 148°C. From some additional experimental work, we have found that the epoxy glass substrate begins to discolor at about 175°C. Thus, in our example, the substrate would not be discolored if the fault current of 10 A is interrupted in about 100 ms. Of course, in an application, one may want to restrict the conductor temperature to much less than 175°C.

It is important to notice that the same conductor temperature would result even if the length, L, of the conductor were only a few inches, since both R_1 and C_1 are proportional to L. Also, for conductor widths other than $W = 7$ mil, one can use the scaling law $C_1 \sim W$ to scale the values of C_1 listed in Table II.

5.4 Some comparative results

Figure 5 compares the transient temperature rises of fine-line printed conductors carrying a current of $I = 2.5$ A on various CP styles. In the steady state, the wire-wrap conductors exhibit the least rise in temperature (21°C); whereas, the conductor on the double-sided CP yields the highest rise in temperature (99°C).

Notice that the relative current-carrying capacity of the various CP styles depends on the duration of current flow. For example, at about $t = 5$ s, the conductor on the double-sided metal board exhibits the highest temperature rise (47°C). Interestingly, this result shows that the thermal conductivity of the dielectric resin of the metal board is somewhat less than the thermal conductivity of epoxy glass. Similarly, Fig. 5 also shows that Milene insulation has a somewhat lower thermal conductivity than Teflon insulation.

In most applications, one is usually interested in the comparative results during the steady state. In this case, the double-sided (epoxy) board is the worst from the point of view of current-carrying capacity.

Finally, notice that in all cases, the conductor temperature rises to a substantial fraction of its final value in the first five seconds.

VI. SUMMARY

This paper presents simple equations, based on the conservation of heat energy, to calculate the transient temperature rise of current-carrying, fine-line (~7 mil) conductors on various styles of circuit packs. The equations depend on various parameters which were determined experimentally. All of the styles of circuit packs in the *BELLPAC* system are included. These styles range from wire wrap to various MLBS.

The maximum steady-state temperature rise of a nicked or constricted current-carrying printed conductor is also treated.

VIII. ACKNOWLEDGMENTS

The author thanks J. J. Waltz, A. P. Irwin, F. A. Scaglione, and J. C. Shank for their valuable support during the course of this study.

REFERENCES

1. W. Aung and A. J. Colucci, unpublished work.
2. W. T. Smith, unpublished work.
3. W. L. Harrod and A. G. Lubowe, "The BELLPAC Modular Electronic Packaging System," B.S.T.J., *58*, No. 10 (December 1979) pp. 2271–88.
4. A. J. Rainal, "Temperature Rise at a Constriction in a Current-Carrying Printed Conductor," B.S.T.J., *55*, No. 2 (February 1976) pp. 233–69.
5. C. W. Jennings, "Electrical Properties of Printed Wiring Boards," Institute of Printed Circuits, IPC-TP-117, September 1976.
6. H. S. Carslaw and J. C. Jaeger, *Conduction of Heat In Solids*, London: Oxford University Press, 1959.
7. Bell Telephone Laboratories Staff, "Integrated Device and Connection Technology", *Physical Design of Electronic Systems*, III, Englewood Cliffs, N. J.: Prentice Hall, 1972.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 60, No. 7, September 1981
Printed in U.S.A.

A Comparative Study of Several Dynamic Time-Warping Algorithms for Connected-Word Recognition

By C. S. MYERS and L. R. RABINER

(Manuscript received February 11, 1981)

Several different algorithms have been proposed for time registering a test pattern and a concatenated (isolated word) sequence of reference patterns for automatic connected-word recognition. These algorithms include the two-level, dynamic programming algorithm, the sampling approach and the level-building approach. In this paper, we discuss the theoretical differences and similarities among the various algorithms. An experimental comparison of these algorithms for a connected-digit recognition task is also given. The comparison shows that for typical applications, the level-building algorithm performs better than either the two-level DP*-matching or the sampling algorithm.*

I. INTRODUCTION

Research in the area of automatic speech recognition has progressed to the point at which a wide variety of isolated word recognition systems have been implemented and used successfully for many applications.[1-9] These systems have been used in such applications as data entry, searching, and sorting; however, use of these systems is restricted by the format of the speech input, i.e. isolated words. For many applications, a connected-word input format would have several advantages. Examples of such applications include:

(*i*) Credit card entry—A sequence of digits (and possibly letters) is required to specify the credit card number.

(*ii*) Directory listing retrieval—A sequence of letters is used to spell the name for which a directory listing is required.

(*iii*) Airline reservations—Sentences based on a restricted vocabulary and a restricted syntax are used to make reservations.

Several techniques for recognizing connected-word sequences from isolated word reference patterns have recently been proposed.[10-15]

In this paper, we compare, both theoretically and experimentally, three algorithms which have been proposed for connected-word recognition. These algorithms are the two-level, dynamic programming matching (TLDPM) approach, the sampling approach, and the level-building (LB) approach.[10-13] Another algorithm of the same general class as the ones considered here has been recently proposed by Bridle; however, it is not considered here.[14]

It should be noted that all of the algorithms which have been proposed for connected-word recognition are loosely related to a general information-theory-based algorithm proposed by Bahl and Jelinek.[15] However, each of these connected-word recognition algorithms make different assumptions, have different implementations, and make specific tradeoffs. Thus, the properties of these algorithms are entirely different from those of Bahl and Jelinek; therefore, they warrant independent study.

In Section II, we review each of the three connected-word recognition algorithms and then provide a theoretical comparison of the general properties. In Section III, we theoretically compare the connected-word recognition algorithms, in Section IV, we describe and give results of several experimental comparisons of these algorithms, and in Section V, we discuss these results and their implications for practical word-recognition systems.

II. CONNECTED-WORD RECOGNITION ALGORITHM

Figure 1 shows the basic structure for all the connected-word recognition algorithms under consideration. The feature extraction is generally similar to that used in most isolated word-recognition systems. Typical feature sets include energy of a set of bandpass filters,[16]

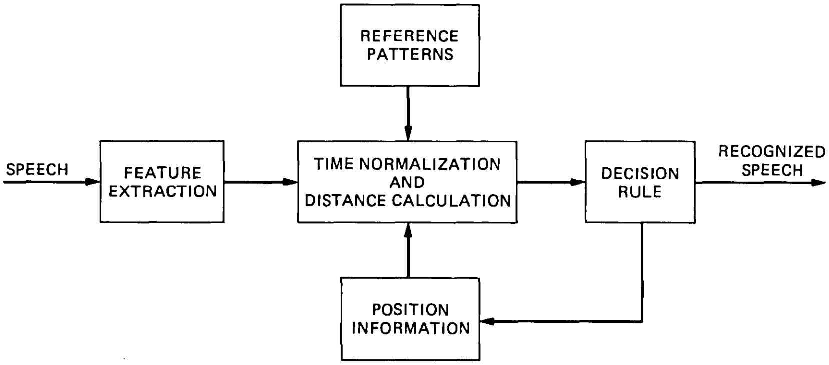


Fig. 1—Block diagram of a generic connected word recognition system.

and LPC coefficients.[14] Following feature extraction, the test utterance, now represented by a sequence of frames of the feature vector, is nonlinearly time-registered, with a set of reference patterns, and a set of distance, or dissimilarity, scores are calculated. Following time registration, a decision rule chooses the sequence of reference patterns which best matches the test utterance. Feedback, in the form of positional information, is used to determine which portion of the test utterance is to be matched to a given reference pattern. It is generally assumed that the test utterance consists of a sequence of words spoken by a *cooperative* user; that is, the words of the string are carefully articulated and spoken in a slow, deliberate manner, but not as isolated words. However, the reference patterns do consist of *isolated* words. The goal of the connected-word recognizer is to find the sequence of concatenated reference patterns, subject to given syntactical constraints, that best matches the test pattern. Among the issues which arise in solving for the best string are the following:

(*i*) How can the reference and test patterns be time-registered?

(*ii*) How can the reference patterns be modified to account for both coarticulation and the natural shortening of words inherent in connected speech?

(*iii*) Is the determination of the best concatenation of reference patterns done in a sequential manner, i.e. one decision at a time, or is some form of backtracking allowed?

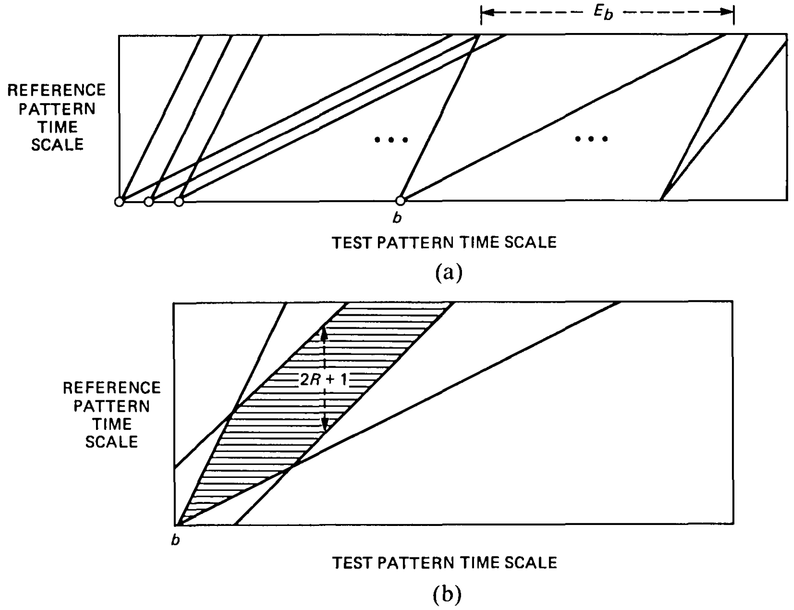
(*iv*) How can alternative matches to the test utterance be generated in addition to the best match?

(*v*) Can syntactical constraints be used explicitly in the recognition stage or is some form of post-processing required?

In the following, we review the three connected-word recognition algorithms and discuss how each of them answers these questions.

2.1 Two-level DP-matching algorithm

The two-level DP-matching (TLDPM) algorithm[10] attempts to find the best concatenation of reference patterns to match a given test pattern (of length M frames) by first determining the optimal reference pattern to match any portion of the test pattern and then attempting to find the optimal way in which to concatenate these pieces. Figure 2a illustrates the first stage of this procedure. For all possible beginning points, b, and all possible references, an isolated word dynamic time warping (DTW) algorithm is used to find the best path to all of the possible ending frames, E_b. Distance scores for the best paths and the reference patterns which generates these best paths are recorded. The region over which a partial isolated word dynamic time warp is examined is shown in Fig. 2b. Here we show a region where the slope of the time warping function is restricted to be between ½ and 2 and



(a)

(b)

Fig. 2—Illustration of the TLDPM algorithm.

where the maximum amount that the time alignment contour may deviate from the line of slope one, which starts at the point $(b, 1)$, is plus or minus R frames.

If we denote the accumulated distance for the vth reference pattern from starting frame b to ending frame e as $\hat{D}(v, b, e)$, then the output of the first stage is the array of $\hat{D}$ values for all v, b, and e combinations. For any set of values of b and e, we can solve for the best reference pattern and distance as

$$D(b, e) = \min_{v} [\hat{D}(v, b, e)], \tag{1a}$$

$$N(b, e) = \operatorname*{argmin}_{v} [\hat{D}(v, b, e)], \tag{1b}$$

where $D(b, e)$ is the minimum accumulated distance from frames b to e of the test pattern, and $N(b, e)$ is the index of the reference pattern giving the minimum distance.

The second stage of the TLDPM algorithm is to determine the best match by piecing together the reference patterns in an optimal manner. This is accomplished again using a dynamic programming algorithm which finds the best concatenation of l reference patterns, ending at frame e of the test pattern by trying all concatenations of a portion of the test pattern ending at frame e and all best reference pattern concatenations of length $l - 1$, i.e.

$$\tilde{D}_l(e) = \min_{1 \le b < e} [D(b, e) + \tilde{D}_{(l-1)}(b - 1)] \tag{2a}$$

$$\tilde{D}_0(0) = 0, \tag{2b}$$

where $\tilde{D}_l(e)$ is the accumulated distance generated by matching the best concatenation of l reference patterns to the portion of the test pattern between frame 1 and frame e and where $D(b, e)$ is the distance associated with the best reference pattern used to match the portion of the test pattern between frame b and frame e. After computation of $\tilde{D}_l(e)$, the best string is recovered by first finding that value of l which minimizes $\tilde{D}_l(M)$ and then tracing back through the sequence of decisions which were used to generate $\tilde{D}_l(M)$.

The TLDPM algorithm made no inherent attempt to modify its reference patterns to account for either coarticulation or word shortening. While word shortening is, in general, compensated directly by the DTW, the use of reference patterns which are inherently longer than the test patterns to which they are to be matched makes this problem much more difficult. In addition, the lack of any boundary modifications for the reference patterns must inherently reduce the potential accuracy of the TLDPM approach.

It is obvious that the TLDPM approach is not a sequential decision method, since no partial match is firmly decided on until the entire second pass of the algorithm is completed. It should also be clear that it is possible to generate not only the best string but the K best strings by simply keeping track of the K best reference patterns which match any portion of the test pattern and also by keeping track of the K best strings for every step of the second pass. Finally, it is clear that, as described, the TLDPM algorithm cannot make use of syntactical constraints directly but must make use of them in a post-processing stage to choose among various candidate strings. That is, syntactical constraints cannot be used to restrict which words of the vocabulary are used for any given beginning and ending point pair, unlike the LB algorithm in which levels correspond to word positions in the string. However, once the beginning and ending point pair distance matrices have been generated, then syntactical constraints can be applied in the second stage of the TLDPM algorithm.

2.2 The sampling approach

Unlike the TLDPM algorithm, the sampling algorithm is more of a sequential decision process. The sampling approach attempts, via a local minimum DTW algorithm,[17] to match a reference pattern to a portion of the test pattern. Unlike the two-level approach of the preceding section, not all portions of the test pattern are tested. Instead, only a small subset of the test pattern is used. The way in which the regions of the test pattern are chosen is as follows. Following the time registration of the reference pattern to one portion of the test pattern, an ending point is implicitly defined by that frame of the test pattern which best matches the end of the reference pattern. After all reference patterns have been tried, the one which gives the best match is chosen as the proper word and the ending point of the test pattern, associated with this reference, is used to hypothesize a beginning region within the test pattern for the next set of references. This procedure is illustrated in Fig. 3. Reference pattern number 1, which is hypothesized to begin somewhere within beginning region 1, is time-registered to the test pattern using a local minimum DTW algorithm. Once the ending point for the best match for Ref. 1 has been determined, a beginning region for the next reference pattern is determined. In general, this beginning region is centered somewhat earlier in the test pattern than the ending position of the previous word. This is done to compensate for the difference in durations between concatenated reference patterns and connected word utterances, and to account for coarticulation between words.

Note that a categorical decision as to the best reference need not be made if the distance scores indicate a high likelihood of confusion. In such cases, when the distance scores for two or more reference patterns are approximately equal, the sampling method keeps tract of all possible strings using the approach described above. Thus, the possi-

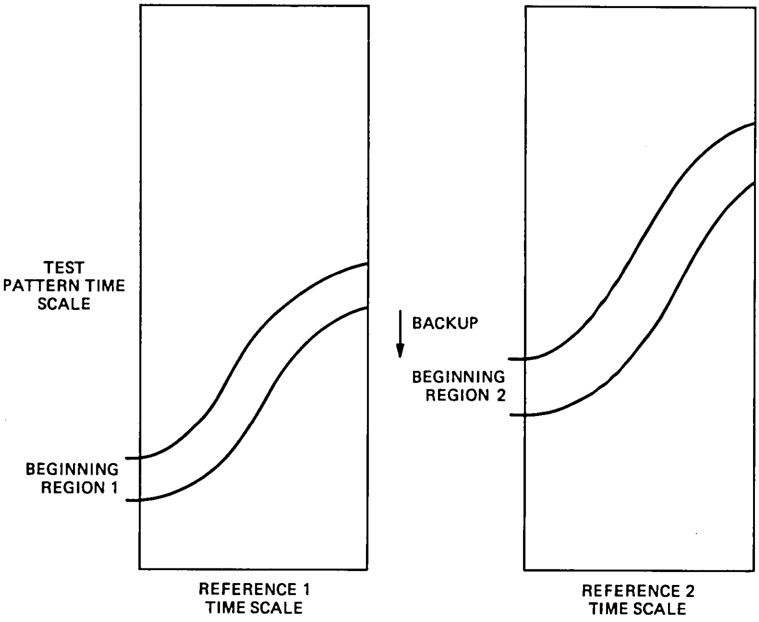


Fig. 3—Illustration of the sampling algorithm.

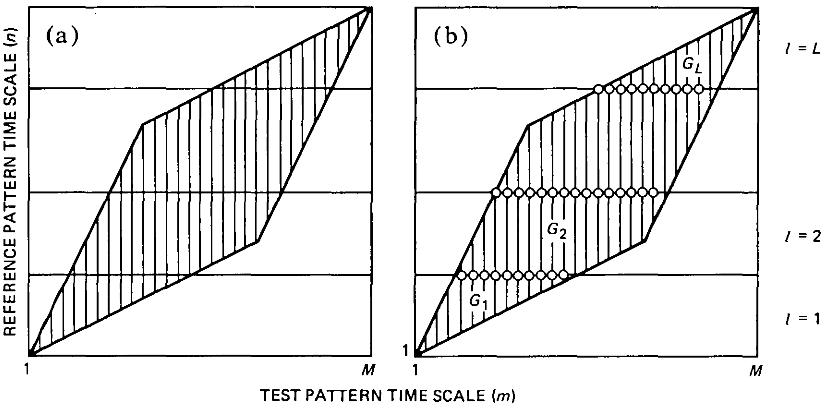


Fig. 4—Two possible implementations of a constrained DTW algorithm.

bility exists of finding a best string at the end of the search, which is different from the best string as the search proceeds from left to right.

In addition to effectively modifying reference patterns by overlapping beginning and ending regions, the sampling algorithm can also use the local minimum DTW algorithm in such a way as to eliminate part of the end of any reference pattern. This is done by not forcing the local minimum DTW algorithm to proceed all the way to the end of the reference pattern, but rather to stop at some frame before the end. Thus, the sampling approach has more inherent flexibility in dealing with modifications to the reference patterns than the TLDPM method.

In contrast to the TLDPM algorithm, the sampling algorithm is able to use syntactical constraints (in the form of a regular grammar) directly, rather than in a post-processing stage. Such a method is described in detail by Levinson and Rosenberg,[12] but the main idea is to trace the graph which represents the grammar simultaneously with matching the reference patterns within the test pattern by keeping track of not only the current state, but also the current ending frame within the test pattern.

2.3 The level-building algorithm

Figure 4 illustrates the basic idea involved in the LB algorithm. Here we show how a constrained endpoint DTW algorithm, in which the slope of the warping function is restricted to be between ½ and 2, is used to find the best alignment between a test pattern and a given concatenated sequence of L reference patterns. In Fig. 4a, we show the computation proceeding in a sequence of vertical strips. Figure 4b shows an alternative way in which the computation may be performed. A set of horizontal lines has been drawn to indicate the boundaries between the different reference patterns in the concatenated reference pattern. We may now compute the optimal, time alignment path in successive levels by using the distances accumulated at the end of one level to initialize the next level. The LB algorithm uses this decomposition to find the optimal sequence of reference patterns by trying all possible reference patterns at any level and recording, for each ending frame at that level, the reference pattern which gave the best distance to that ending frame and a pointer back to the previous levels. These minimum distances are used to initialize the following level. After the final level has been examined, the optimal path is recovered by tracing back along the chain of pointers.

We have demonstrated that the algorithm solves exactly the same problem as the TLDPM algorithm.[13] In addition, we have shown how to modify the LB algorithm via a set of parameters to give it more flexibility. These parameters, as shown in Fig. 5, are as follows:

(*i*) δ_{R_1} –Region of uncertainty at the beginning of the reference pattern.

(*ii*) δ_{R_2} –Region of uncertainty at the end of the reference pattern.

(*iii*) δ_{END} –Region of uncertainty at the end of the test pattern.

(*iv*) M_T –Multiplier used to reduce the size of the beginning region for any level.

(*v*) ϵ –Parameter used to restrict the size of any vertical strip.

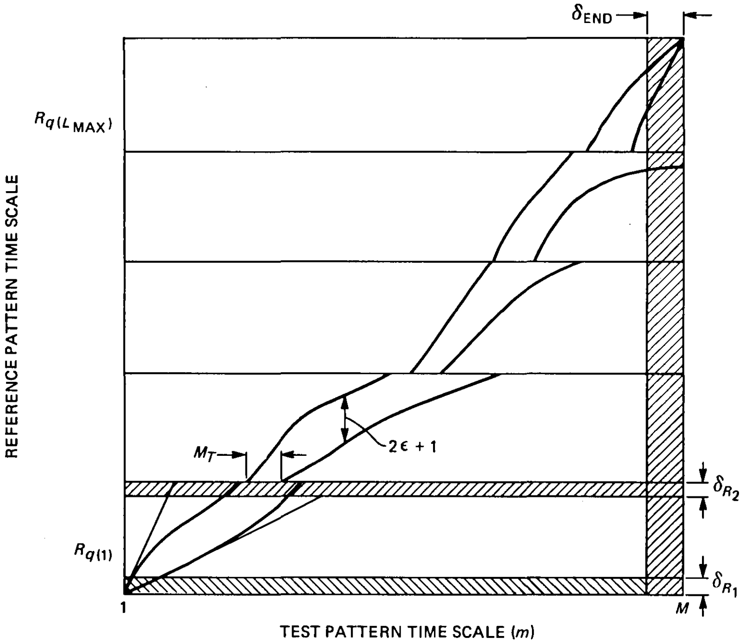


Fig. 5—Illustrations of the parameters of the LB algorithm.

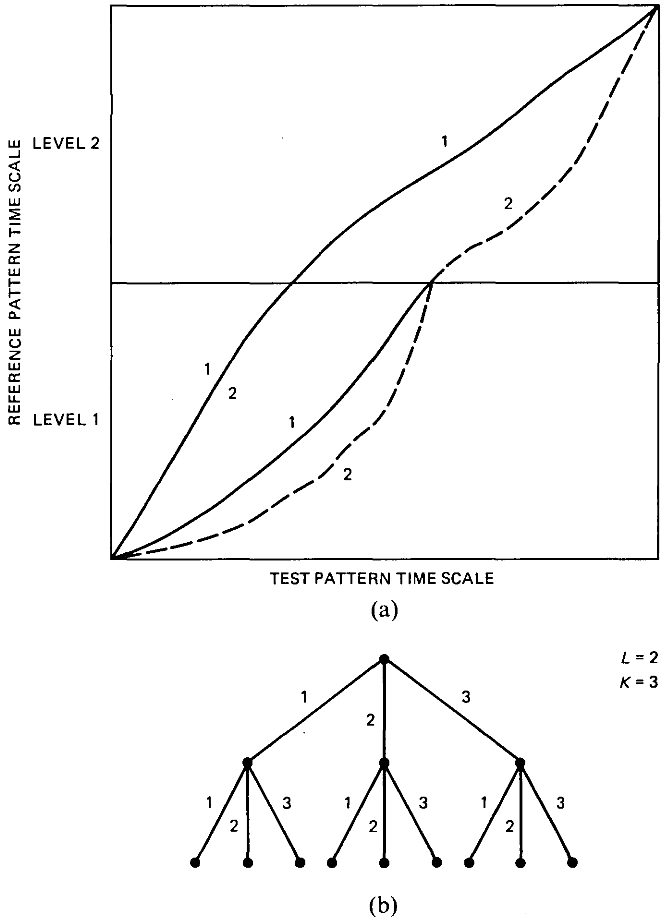


Fig. 6—Generation of multiple candidate strings for the LB algorithm.

The effects of these parameters are shown in Fig. 5. We observe that δ_{R_1} and δ_{R_2} define regions, at the beginning and end of each reference pattern, in which the local path may begin or end. In this manner, some of the gross features of coarticulation and length reduction present in a connected-word utterance may be accounted for. Thus, the LB algorithm has some of the flexibility inherent in the sampling algorithm.

In addition to modifying templates, the algorithm has the advantage of not being a sequential decision process and, thus, has the ability to recover from mistakes as in the TLDPM algorithm.

One important shortcoming of the LB algorithm is its ability to generate alternative candidates. The method by which multiple candidates are generated is to record not only the best candidate to each ending frame of each level, but to record the K best candidates and then to allow substitutions in the traceback. This method is illustrated in Fig. 6a for $K = 2$ candidates and for $L = 2$ levels. The solid paths represent the best paths from the beginning of the level to that particular ending point and the dashed paths represent the second best path to that particular ending point (using a different reference than the one used in the best path). We see that for $K = 2$ and $L = 2$ we may generate four different candidate strings and, in general, may have K^L different candidate strings. Such a situation may be represented by a tree with a branching factor of K and a depth of L as shown in Fig. 6b for $K = 3, L = 2$. Generation of the possible candidate strings is simply a tree-searching problem, and it is possible to reduce the amount of traceback by pruning the tree. To prune the tree we may take advantage of the fact that the kth best path at any node in the tree represents a string whose distance is always larger than the $(k - 1)$st best path to that node. The difficulty with such a scheme is that, unlike the TLDPM algorithm, we are not guaranteed of finding even the true second best path. Figure 7 illustrates this problem. Here the best path is given by string AA and an alternative is given by string CA. Another path, shown by string BA, must have a larger distance than string AA, but may have a smaller distance than string CA. However, the path BA will not be recorded because, at the second

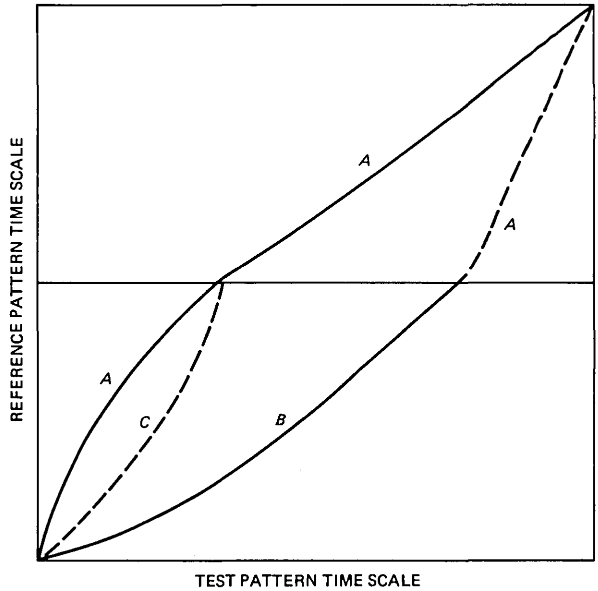


Fig. 7—Illustration of the failure of the LB algorithm to find all good candidate strings.

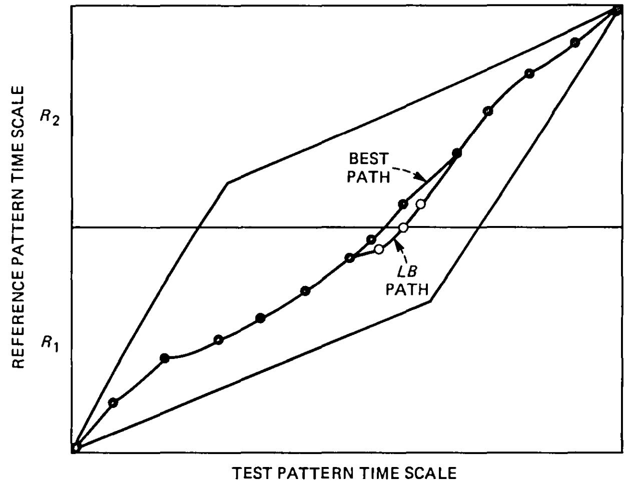


Fig. 8—Illustration of the path differences between the LB algorithm and a single time warp to the concatenated sequence of reference patterns.

level, only the *best* path using reference A is recorded. While it would be possible to record this second best path, the computational problems which would have to be overcome seem much too burdensome.

Another small theoretical difference between both the TLDPM and the LB solutions, and the exact best solution for the sequence of concatenated reference patterns that best matches the test pattern, is illustrated in Fig. 8. Here we show a sequence of two reference patterns, R_1 and R_2, and the best path for warping the super reference pattern formed by concatenating R_1 and R_2 to T, as well as the LB path. Since the LB path is constrained to end at the end of each reference pattern (level), then a path point is constrained to occur at the end of each reference pattern. This need not be the case for the optimum concatenation and, thus, a small difference can exist in the solutions. In practice, this effect does not occur (unless $\delta_{R_1} = \delta_{R_2} = 0$) since the δ_{R_1}, δ_{R_2} parameters allow paths to skip frames of the reference patterns.

A final consideration involving the LB algorithm is the use of syntactical constraints. Since all good candidates may not be generated by the LB algorithm, it is important that syntactical constraints be easily incorporated into its structure. Fortunately, this is possible and has been described in detail by Myers and Levinson.[18] The basic principle is to build up the results by states of a finite-state automata, rather than by levels, and to use transitions of the finite state automata to guide the recognition process.

In Section III, we summarize the qualitative features of the different connected-word recognition algorithms and also give a quantitative evaluation of both their space and time complexities.

III. THEORETICAL COMPARISON OF THE CONNECTED-WORD RECOGNITION ALGORITHMS

In Table I we summarize our qualitative evaluation of the different connected-word recognition algorithms. Both the sampling and LB algorithms have the ability to modify their reference patterns, although in different ways. The sampling approach allows both the removal of some frames from the reference pattern and the overlap of reference patterns, while the LB approach simply allows the removal of some frames from the reference patterns. Thus, we may conclude that the LB and the sampling algorithms will be able to more easily match test patterns which contain large amounts of coarticulation or length shortening than the TLDPM algorithm.

We have shown that the sampling algorithm is a sequential decision process, while both the TLDPM and the LB algorithms have some form of backtracking. Thus, we expect better performance from the TLDPM and the LB algorithms in those cases in which the sampling procedure may get lost.

In addition to being able to avoid getting lost, the TLDPM algorithm is the only algorithm that is capable of generating alternative candidates which are exactly correct. Thus, this algorithm should perform the best in cases in which many potential candidates are desirable. Unfortunately, as the final entry shows, the TLDPM algorithm is not well suited for syntactical analyses; therefore, it is more difficult to implement constraints using this algorithm.*

Table I—Qualitative comparison of connected-word recognition algorithms

Algorithm	Modification of References	Sequential Decision	Multiple Candidates	Syntax Constraints
Two-level DP-matching	No	No	Exact	No
Sampling	Yes	Yes	Heuristic	Yes
Level building	Yes	No	Heuristic	Yes

3.1 Computational comparison

In comparing both the time and space complexity of the three connected-word recognition algorithms, we shall assume that we are given a test utterance of length M frames, a vocabulary containing V words, a set of reference patterns with average length $\bar{N}$ (frames), with maximum time alignment deviation of $\pm R$ frames. For purpose of

* Note that it *is* possible to implement both syntactical constraints and reference pattern modifications directly in the TLDPM algorithm without changing the fundamental ideas contained in the algorithm. For purposes of our comparison, however, we have considered the TLDPM algorithm only as it was originally described.

comparison, our measure of time will be the number of times that a frame of the test utterance is compared to a frame of any reference pattern. For the TLDPM algorithm, the number of comparisons is given by

$$NC_{\text{TLDPM}} = V \cdot M \cdot \bar{N} \cdot (2R + 1), \tag{3}$$

since there is an isolated word time-warp of size $\bar{N} \cdot (2R + 1)$ for all V possible words, and for each of the M test frames.

For the sampling algorithm, the number of comparisons is given by

$$NC_s = V \cdot L \cdot \bar{N} \cdot (2\bar{\epsilon} + 1) \cdot \bar{\gamma}, \tag{4}$$

where L is the actual number of words in the test utterance, $\bar{\epsilon}$ is the range for the local minimum DTW algorithm, and $\bar{\gamma}$ is the average number of candidate strings which are retained. (Rabiner and Schmidt found that $\bar{\gamma}$ was, on average, between 1 and 2).[11] Data in this figure result from one local minimum time-warp of size $\bar{N} \cdot (2\bar{\epsilon} + 1)$ for each possible reference, for each word of the test pattern and for each candidate string.

For the LB algorithm, in which the slope of the warping function is restricted to be between ½ and 2 (as shown in Fig. 4), the number of comparisons is given by

$$NC_{\text{LB}} = V \cdot L_{\text{MAX}} \cdot M \cdot \bar{N}/3, \tag{5}$$

because the size of parallelogram in Fig. 4 is about $M \cdot \bar{N} \cdot L_{\text{MAX}}/3$, and because all V references must be used at each level.

Finally, by incorporating the range reduction techniques described by Myers and Rabiner,[13] the number of comparisons required for the reduced LB technique becomes

$$NC_{\text{LBR}} = V \cdot L_{\text{MAX}} \cdot \bar{N} \cdot (2\epsilon + 1), \tag{6}$$

since the reduced LB technique uses only a single local minimum time warp per level.

Table IIa summarizes the computational aspects of these connected-word DTW algorithms. The row-labelled number of basic time warps refers to the number of times that a basic DTW algorithm is applied. The size of the time warp is the average size of these basic time warps, and the total is the number of comparisons given in eqs. (3) to (6).

Table IIb gives a numerical comparison of the computation required by the various connected-word recognition algorithms for the case of

$$L_{\text{MAX}} = 5,\ V = 10,\ M = 120,\ \bar{N} = 35,$$

$$\bar{\epsilon} = 8,\ \epsilon = 12,\ R = 12,\ \bar{\gamma} = 1.5,\ L = 4.$$

Note that the total computation required by the TLDPM algorithm is 15 times that of the LB algorithm and 30 times that of either the

Table IIa—Computational comparisons of connected-word DTW algorithm

	Level Building	Two-Level DP Matching	Sampling	Reduced Level Building
Number of basic time warps	$L_{\text{MAX}} \cdot V$	$M \cdot V$	$L \cdot V \cdot \bar{\gamma}$	$L \cdot V$
Size of time warps	$\bar{N} \cdot M/3$	$\bar{N} \cdot (2R + 1)$	$\bar{N} \cdot (2\bar{\epsilon} + 1)$	$\bar{N} \cdot (2\epsilon + 1)$
Total computation for distance	$L_{\text{MAX}} \cdot V \cdot \bar{N} \cdot M/3$	$M \cdot V \cdot \bar{N} \cdot (2R + 1)$	$L \cdot V \cdot \bar{\gamma} \cdot \bar{N}(2\bar{\epsilon} + 1)$	$L \cdot V \cdot \bar{N}(2\epsilon + 1)$
Storage	$3 \cdot M \cdot L_{\text{MAX}} \cdot K$	$2 \cdot M \cdot (2R + 1) \cdot K$	0	$3 \cdot M \cdot L_{\text{MAX}} \cdot K$

Table IIb—Typical computational requirements for the case $L_{\text{MAX}} = 5$, $V = 10$, $M = 120$, $\bar{N} = 35$, $\bar{\epsilon} = 8$, $\epsilon = 12$, $\gamma = 1.5$, $L = 4$, $K = 2$, $R = 12$

	Level Building	Two-Level DP Matching	Sampling	Reduced Level Building
Number of basic time warps	50	1200	60	40
Size of time warps	1400	875	595	875
Total computation for distances	70,000	1,050,000	35,700	35,000
Storage	3600	12000	0	3600

reduced LB or the sampling algorithm. The efficiency of the LB algorithm derives partly from the fact that optimal paths to a range of ending frames for a range of starting frames are found simultaneously.

3.2 Storage comparison

In comparing the storage required by the various connected-word recognition algorithms, we refer only to that storage which varies among the different algorithms. Hence, storage for reference patterns, and storage needed by all basic DTW algorithms is not considered here. The storage for the TLDPM algorithm is given by

$$S_{\mathrm{TLDPM}} = 2 \cdot M \cdot (2R + 1) \cdot K. \qquad (7)$$

Storage is required for the matrices of the K best distances and their associated words, for each of the $M \cdot (2R + 1)$ pairs of beginning and ending frames.

The storage for the sampling algorithm is so small (less than 100) as to be negligible. The storage for the LB algorithm is given by

$$S_{\mathrm{LB}} = 3 \cdot M \cdot L_{\mathrm{MAX}} \cdot K \qquad (8)$$

since, at the end of each level, it is necessary to store a distance, an associated word, and a pointer to the previous level for each of the K possible candidates. (Clearly, the factor M in eq. (8) can be substantially reduced if storage of distances, words, and back pointers is used only for finite distance ending frames.)

The storage requirements for the four algorithms are summarized in the last row of Table IIa, and a numerical example ($K = 2$) is given in Table IIb. Note that both the LB and the TLDPM algorithm require a significant amount of storage, but that the storage required by the LB algorithm is only a third of that required by the TLDPM algorithm.

We have seen that both the LB and the sampling algorithm are computationally simpler than the TLDPM, but that the TLDPM algorithm has the potential to generate better candidates for a given connected-word recognition task. Thus, certain trade-offs exist and some important questions must now be considered. Among them are the following:

(*i*) Is the ability to modify the reference patterns a useful feature for a connected-word recognition algorithm?

(*ii*) Is the ability to backtrack a useful feature, or is a sequential decision process sufficient?

(*iii*) Are the additional computational costs of the TLDPM worth the increased number of good candidate strings?

In Section IV we give the results of several experimental studies designed to answer these questions.

IV. EXPERIMENTAL COMPARISONS OF CONNECTED-WORD RECOGNITION ALGORITHMS

To answer the questions at the end of Section III, each of the three connected-word recognition algorithms was simulated and tested on a common data base of 480 connected digit strings. The data base consisted of 80 strings of variable length (from two to three connected digits) spoken by each of six speakers (three male, three female). The strings were carefully articulated, and spoken in a slow, deliberate manner. Care was taken to guarantee an equal number of occurrences of all length strings, (from two to five digits), and an equal number of occurrences of all digits within the strings. This data base was used previously to evaluate the sampling method and the LB approach.[11,19]

Since results had been obtained previously on the sampling and LB algorithms, the TLDPM algorithm was all that had to be simulated and tested. For consistency, the LPC feature set used in the sampling and LB systems was used in the TLDPM system as the basic analysis features. The only variable parameter in the basic two-level DP warp method is the time adjustment parameter R. Sakoe indicated that a value of $R = 0.4\bar{N}$ would be appropriate where $\bar{N}$ is the average reference length.[10] Since $\bar{N} = 40$, this indicated that a value of $R = 16$ was required. The first experiment varied R from 8 to 99 to see its effect on recognition accuracy for a speaker-trained system. The results of this experiment (i.e., string error rate versus R) are given in Fig. 9. The results shown indicate that the larger the value of R, the lower the error rate, and that a flattening on the error curve occurs for values of $R \geq 20$.

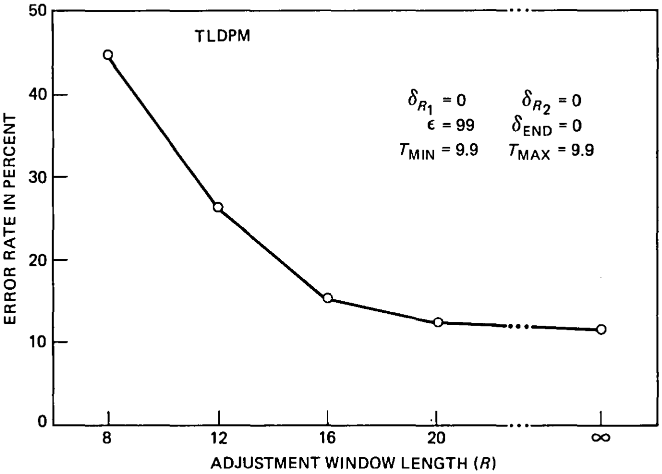


Fig. 9—String error rate as a function of the time shift parameter R of the TLDPM method for connecting digit strings.

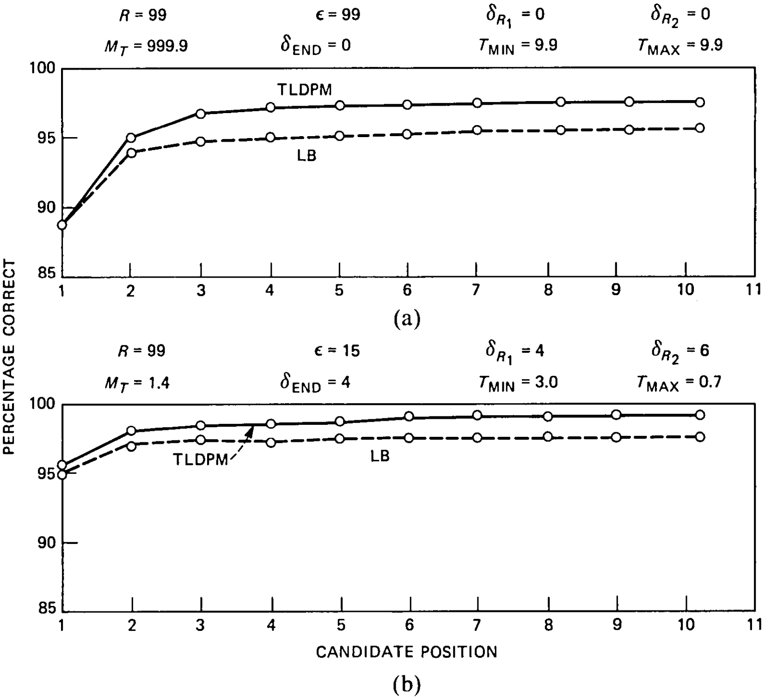


Fig. 10—Percentage correct strings as a function of candidate position for the TLDPM method and the LB method for (a) no reference modification and for (b) reference modification.

The next experiment made a direct comparison between the TLDPM method and the LB algorithm by setting the flexible LB parameters to values appropriate for the full range (ϵ = 99, M_T = 99.9, δ_{END} = 0, T_{MIN} = 9.9, T_{MAX} = 9.9, T_{MIN} and T_{MAX} specify distance thresholds used to reduce computation), and for no reduction in template lengths $\delta_{R_1} = \delta_{R_2} = 0$).[13] The $K = 2$ best candidates were recorded at each level and the 10 best strings (of any length) were found from the LB output. Similarly, for the two-level DP warp, the 10 best strings were found directly. Figure 10a shows the percentage of strings correct as a function of candidate position in the list for both the TLDPM method (solid curve) and the LB method (dashed curve). Again, a speaker-trained system was used. The results show the same string error rate (11.5 percent) on the top candidate for both methods as anticipated from the earlier discussion. However, the results also show a 1 to 2 percent higher string accuracy for the TLDPM method for candidate positions 2 to 10, indicating the potential improvements obtained by keeping distances for all posible beginning and ending frames.

Since the overall performance of the LB method was significantly better, using the best values of the LB parameters, rather than those shown in Fig. 10a, another experiment was run to compare recognition accuracy versus candidate position for both methods at the speaker-trained operating point ($R = 99$, $\epsilon = 99$, $M_T = 1.4$, $\delta_{\text{END}} = 4$, $T_{\text{MIN}} = 3.0$, $T_{\text{MAX}} = 0.7$, $\delta_{R_1} = 4$, $\delta_{R_2} = 6$). It should be noted that in this case, the TLDPM algorithm was modified from its original specifications (to a form similar to the UE2-1 algorithm[17]) to incorporate the parameters of the LB algorithm. These results are plotted in Fig. 10b. For the best candidate, the string error rates were 4.6 percent for the TLDPM method, and 4.8 percent for the LB method. For candidate positions 2–10, the TLDPM method has from 1 to 1.7 percent higher accuracy than the LB method. These results again indicate the small, but consistent improvement obtained by the TLDPM method, at least for alternative best string candidates. Whether or not this improved accuracy on higher candidate positions can be used depends strongly on the task, and the types of errors that occurred. Clearly, errors in string length can often be simply corrected, whereas errors of the same string length are generally difficult to detect and correct.

A summary of the resulting string error rates on the top recognition candidate for all three connected-word recognizers is given in Table III. For the sampling and LB methods, results are also given for performance in a speaker-independent mode (using 12 isolated speaker-independent templates per digit). No results are given for the TLDPM system as a speaker-independent recognizer because of the inordinate amount of running time (about 4 to 5 hours/string) required to evaluate its performance.

4.1 *Performance of the LB method for other applications*

The LB method of connected-word recognition was also applied to the problem of recognizing names from a given directory from spoken, spelled input. Since the vocabulary here is letters of the alphabet, and because of the high degree of confusability among several of the letters (e.g., *B*, *D*, *P*, *V*, etc.), a somewhat different structure was used for recognizing the names.[9] First, a name class is found, and then all names (if any) within the name class are tested and a name score is stored.

Table III—String error rates on connected digits

	Word Recognition Method			
	TLDPM (original)	TLDPM (modified)	Sampling	LB (reduced)
Speaker-trained	11.5%	4.6%	6.7%	4.8%
Speaker-independent	-	-	9.0%	4.6%

Table IV—Percentage names correct using the LB method

	Talking Speed	
	Deliberate	Normal
Speaker-trained	99%	90.5%
Speaker-independent	96%	87.5%

The name with the smallest name score is chosen as the recognized name.

To evaluate the performance of this system, four speakers (two male, two female) spoke 50 randomly selected names each as a connected sequence of letters of the last name, followed by a pause, followed by the initials. Each speaker spoke the name list two times. The first time, they spoke the names in a deliberate, carefully articulated manner, the second time, in a normal manner. The overall name accuracy achieved using the LB system is given in Table IV for both a speaker-trained and a speaker-independent implementation. Name accuracies of 96 to 99 percent were obtained for the deliberately spoken names, and of 87.5 to 90.5 percent for normally spoken names.

In addition to directory assistance, the LB algorithm has also been applied to the problem of sentence recognition for an airlines reservation system.[18] In this experiment, four speakers (two male and two female) spoke 50 sentences each. These sentences were formed from a 127-word vocabulary using a regular grammar. The grammar consisted of 144 states and 450 transitions. The sentences were spoken in a normal manner. Recognition was performed in a speaker-dependent manner. Sentence accuracies of 89 percent and word accuracies of 94 percent were obtained. These experiments on directory assistance and sentence recognition demonstrate that the LB algorithm has widespread applicability.

V. DISCUSSION AND SUMMARY

The answers to two of the three questions at the end of Section III are clear. The gain in accuracy obtained by modifying the reference pattern is quite large for the connected-digit sequences (e.g., the string error rate falls from 11.5 percent to 4.6 percent for the TLDPM method). Similar improvements have previously been noted for the LB algorithm.[19] Thus, it is clear that the modified reference patterns lead to improved overall performance. However, the results on recognizing spelled names (see Table IV) indicate clearly that as the rate of talking goes up, the performance of the system becomes dramatically worse. At this point, the basic model assumptions begin to fall apart and no simple modification of the reference patterns is adequate.

The answer to the second question concerning the value of a backtracking versus a sequential approach can be obtained from Table III. By comparing performance of the sampling approach (with sequential decisions) to the LB or TLDPM results (with backtracking), we see a loss in string accuracy of 2 to 4.4 percent using the sequential approach. Since there is essentially no computational gain which accompanies this loss in accuracy, it seems clear that a backtracking approach is superior to a sequential decision method.

The final question in section IV—whether the improved accuracy in higher candidate positions of the TLDPM method justifies the greatly increased computational load—is hard to answer. Since the recognition accuracy on the best string is the same, and since a cost factor in computation of about 40-to-1 is incurred, it would appear that the increased accuracy is not sufficient to justify the increased computation. However, there may exist tasks with constrained syntax such that the improved accuracy does justify the costs.

The overall conclusion is that both the LB and TLDPM methods provide high accuracy in recognizing strings of connected words. Moreover, since the computation of the LB method is considerably less than the computation of the TLDPM, this method is to be preferred for most applications.

REFERENCES

1. T. B. Martin, "Practical Applications of Voice Input to Machines," Proc. IEEE, *64* (April 1976), pp. 487-501.
2. M. B. Herscher and R. B. Cox, "Source Data Entry Using Voice Input," Conf. Record 1976 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (April 1976), pp. 190-3.
3. S. L. Moshier, "Talker Independent Speech Recognition in Commercial Environments," Speech Comm. Papers, Acoustics Soc. of Amer., Paper YY12, J. J. Wolf and D. H. Klatt, Eds., June 1979, pp. 551-3.
4. F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans. on Acoustics, Speech, and Signal Processing, *ASSP-23,* No. 1 (February 1975), pp. 67-72.
5. A. E. Rosenberg and F. Itakura, "Evaluation of an Automatic Word Recognition System Over Dialed-Up Telephone Lines," J. Acoust. Soc. Am., *60,* Suppl. 1 (November 1976), pp. S12 (Abstract).
6. S. E. Levinson et al., "Interactive Clustering Techniques for Selecting Speaker-Independent Reference Templates for Isolated Word Recognition," IEEE Trans. on Acoustics, Speech, and Signal Processings, *ASSP-27,* No. 3 (April 1979), pp. 134-41.
7. L. R. Rabiner et al., "Speaker-Independent Recognition of Isolated Words Using Clustering Techniques," IEEE Trans. on Acoustics, Speech, and Signal Processing, *ASSP-27,* No. 4 (August 1979), pp. 236-349.
8. L. R. Rabiner, J. G. Wilpon, and A. E. Rosenberg, "A Voice Controlled Repertory Dialer System," B.S.T.J., *59* (September 1980), pp. 1153-63.
9. B. Aldefeld et al., "Automated Directory Listing Retrieval System Based on Isolated Word Recognition," IEEE Proc., *68,* No. 10 (November 1980), pp. 1364-79.
10. H. Sakoe, "Two Level DP-Matching-A Dynamic Programming Based Pattern Matching Algorithm for Connected Word Recognition," IEEE Trans. on Acoustics, Speech, and Signal Processing, *ASSP-27,* No. 6 (December 1979), pp. 588-95.
11. L. R. Rabiner and C. E. Schmidt, "Application of Dynamic Time Warping to Connected Digit Recognition," IEEE Trans. on Acoustics, Speech, and Signal Processing, *ASSP-28,* No. 4 (August 1980), pp. 377-88.

12. S. E. Levinson and A. E. Rosenberg, "A New System for Continuous Speech Recognition—Preliminary Results," Proc. Int. Conf. Acoustics, Speech, and Signal Processing (April 1979), pp. 239–44.
13. C. S. Myers and L. R. Rabiner, "A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition," IEEE Trans. on Acoustics, Speech, and Signal Processing, *ASSP-29*, No. 2 (April, 1981), pp. 284–97.
14. J. S. Bridle and M. D. Brown, "Connected Word Recognition Using Whole Word Templates," Proc. of the Institute of Acoustics, Autumn Conf., 1979.
15. L. R. Bahl and F. Jelinek, "Decoding for Channels with Insertions, Deletions, and Substitutions with Applications to Speech Recognition," IEEE Trans. on Info. Theory, *IT-21*, No. 4 (July 1975), pp. 404–11.
16. H. F. Silverman and N. R. Dixon, "A Comparison of Several Speech-Spectra Classification Methods," IEEE Trans. on Acoustics, Speech, and Signal Processing, *ASSP-24*, No. 4 (August 1976), pp. 289–95.
17. L. R. Rabiner, A. E. Rosenberg, and S. E. Levinson, "Considerations in Dynamic Time Warping for Discrete Word Recognition." IEEE Trans. on Acoustics, Speech, and Signal Processing, *ASSP-26*, No. 6 (December 1978), pp. 575–82.
18. C. S. Myers, and S. E. Levinson, "Sentence Recognition Using a Level Building Dynamic Time Warping Algorithm," Proc. Int. Conf. Acoustics, Speech, and Signaling Processing, 1981, April 1, 1981, pp. 956-9.
19. C. S. Myers and L. R. Rabiner, "Connected Digit Recognition Using a Level Building DTW Algorithm," IEEE Trans. on Acoustics, Speech, and Signal Processing, *ASSP-29*, No. 3 (June 1981), pp. 351-63.

Copyright © 1981 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 60, No. 7, September 1981
Printed in U.S.A.

A Comparison of Three Speech Coders to be Implemented on the Digital Signal Processor

By R. V. COX

(Manuscript received June 24, 1980)

The recently developed digital signal processor is a device used for implementing low- to medium-complexity speech coders. It is currently being used in implementing adaptive differential pulse-code modulation (ADPCM) coding, two-band sub-band coding, and four-band sub-band coding. This study was performed to determine optimal parameter values for the two sub-band coders in preparation for their implementation on the digital signal processor and to determine their performance relative to ADPCM. (The actual implementation of the ADPCM and two-band sub-band algorithms are discussed in other papers in Part 2 of this issue of the Bell System Technical Journal.*) Performance was judged on the basis of segmental signal-to-noise ratio and a forced-choice, subjective comparison test of the coders. All three coders were simulated at bit rates of 16, 20, 24, 28, and 32 kb/s. The simulations were performed on a laboratory computer.*

I. INTRODUCTION

The recently developed DSP is a device for implementing low-to medium-complexity speech coders. Three coders are currently being implemented. The simplest coder is adaptive differential pulse-code modulation (ADPCM) and is discussed in Ref. 1. The other two are in the sub-band coder (SBC) family. Of these, the simpler one is two-band sub-band coding (2B-SBC), featuring quadrature mirror filtering and two equal bands. It is discussed in Ref. 2. The other coder—the most complicated—is four-band sub-band coding (4B-SBC), featuring four equal bands. Its implementation is still in progress.

This report discusses the initial design parameters for the latter two coders and the relative performance of all three. Segmental signal-to-noise ratio (SNR) measurements were made on all three coders via computer simulation at five different bit rates, 16, 20, 24, 28, and 32 kb/s. In addition, 12 subjects ranked the coders in a comparison test. The simulations reported here were carried out on a laboratory computer as preparation for the implementation of the SBC coders on the DSP.

Section II reviews the design of ADPCM and discusses the design of the two sub-band coders. Section III discusses the results of the subjective testing experiment, and Section IV gives the conclusions of this study.

II. DESIGN OF THE CODERS

2.1 Design of ADPCM

The ADPCM design simulated here is based on the design of Cummiskey et al.[3] A block diagram of the ADPCM coder described below is shown in Fig. 1. The most significant change from the design in Ref. 3, is that only two multiplier values are used in changing the step-size, regardless of the bit rate. This is based on the ADPCM implemented by Johnston and Goodman.[4] This version of ADPCM has already been

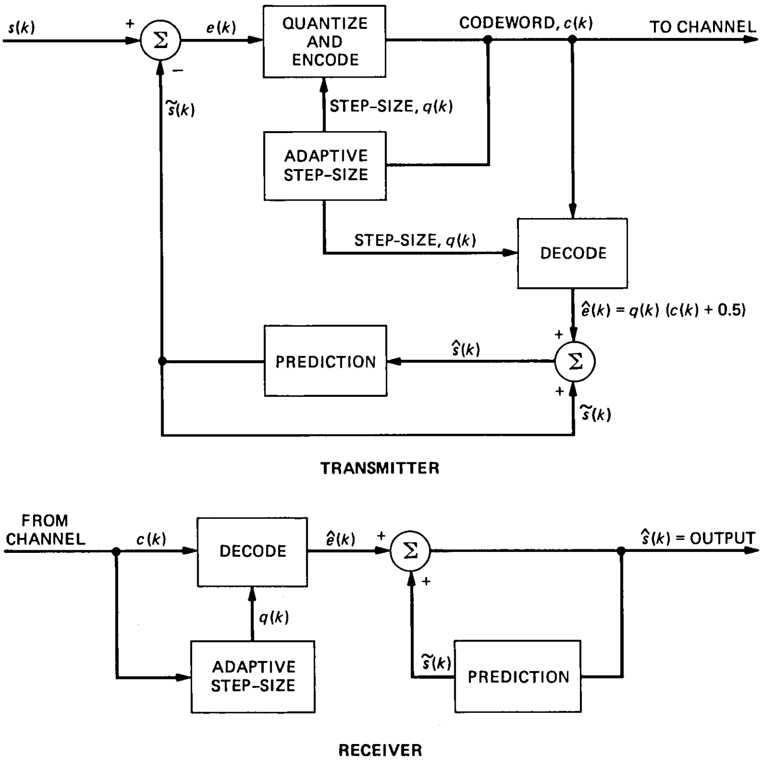


Fig. 1—Adaptive differential pulse-code-modulation coder used in simulation.

implemented on the DSP and is described in Ref. 1. The ADPCM design was also used for quantizing the sub-band signals in the other coders.

To simulate 20 kb/s and 28 kb/s ADPCM, alternating quantizers were used. For 20 kb/s, the two quantizers used are for 2 and 3 bits. Since the step-size is adapted based only on the most significant magnitude bit, the same step-size adaptation algorithm is used for all samples. The ratio of 2- to 3-bit quantizer step-size is held constant. This requires one additional multiplication to convert from the 2- to 3-bit step-size.

2.2 Two-band SBC

All sub-band coders are made from a few fundamental building blocks. The first is linear filtering to divide the signal into two or more sub-bands. These sub-bands can then be decimated to a lower sampling rate than the original signal. Some form of quantization must be used to encode and quantize each band. Interpolation and additional linear filtering is used to bring each band back to the original sampling rate and to its original space in the frequency spectrum. At this point, they can be added together to produce an output signal.

The quadrature mirror filtering technique is fairly well known for its use with sub-band coders. Each pair of quadrature mirror filters (QMFs) produces two sub-bands of equal width in frequency. Johnson has compiled a collection of different length QMFs.[5] The possible quantizers which can be used are adaptive delta modulation (ADM), ADPCM, and adaptive pulse-code modulation (APCM). Each of these techniques is fairly well known and understood. Likewise interpolation and decimation are also well understood. So what remains is the task of combining these building blocks in such a way as to fit on the DSP and, also, give the best possible performance. One of the tasks of this study was to choose good candidates for implementation.

The 2B-SBC design is based on the 2B-SBC commentary grade coder of Johnston and Crochiere.[6] That coder was developed with the object of maintaining a high-quality AM radio signal. Its parameters were tuned to music rather than to speech. This section describes parameters for a speech bandwidth version. There are five possible bit rates envisioned. For a more detailed discussion of sub-band coding in general and the exact implementation of this coder refer to Ref. 2.

Figure 2 is a block diagram of the 2B-SBC. The input speech has been band-limited from 200 to 3200 Hz by a sharp bandpass filter and sampled at 8000 Hz. A 32-tap QMF designed by Johnston is used for separating the digitized speech into the two sub-bands.[5] After 2-to-1 decimation on both bands, we found average correlations for speech of 0.7 and −0.45 for the low and high bands, respectively.

The two bands are then coded using either ADPCM or ADM. The ADM

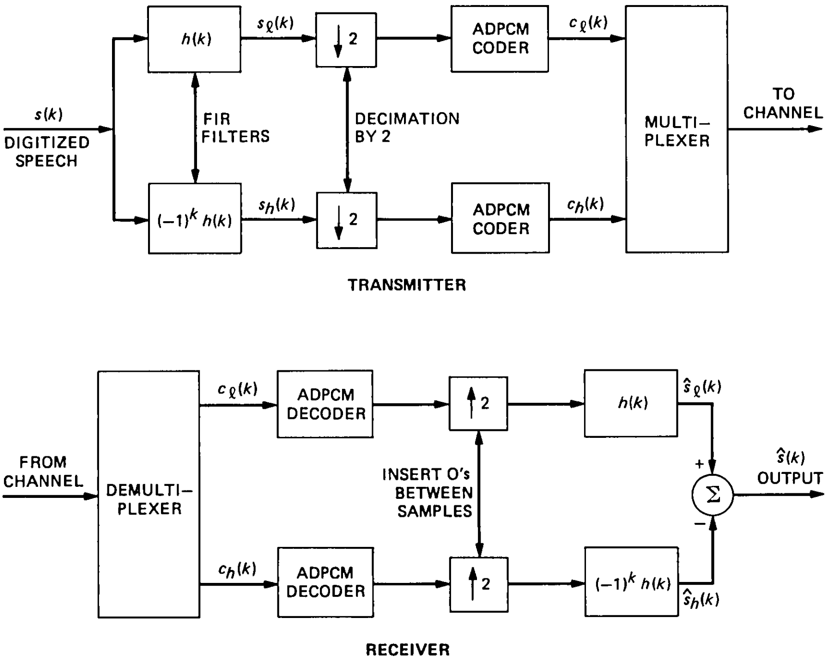


Fig. 2—Two-band sub-band coder used in simulation.

is used only for the higher band at low bit rates. It is based on the ADM of Jayant.[7] The prediction coefficients used for these coders were the correlation values mentioned above. Since the high band has a negative correlation, it was frequency inverted before quantization by the ADM, because this ADM requires a positive correlation for its adaptation mechanism to work properly. Since frequency inversion just means changing the sign of every other sample, this is a very mınor operation.

The next step was to determine optimal bit allocations for the low and high bands. After experimenting with different bit allocations and evaluating them on the basis of segmental SNR measurements and informal listening, the following bit allocations were adopted for the five bit rates:

16 kb/s low: 3 bits high: 1 bit (ADM)
20 kb/s low: 4 bits high: 1 bit (ADM)
24 kb/s low: 5 bits high: 1 bit (ADM)
28 kb/s low: 5 bits high: 2 bits
32 kb/s low: 5 bits high: 3 bits.

Some alternative designs were very close. For instance, a (4, 2) allocation for 24 kb/s is almost as good as (5, 1) for the speech it was tested on. Perhaps if the speech were less sharply bandpass-filtered and if there were more high-frequency content (such as in telephone speech) the better allocation would be (4, 2) for 24 kb/s.

2.3 Four-band SBC design

The 4B-SBC design described here is new, although it is a logical extension of the 2B-SBC mentioned above. It starts with the same two sub-bands as the two-band design. Both of these bands are then divided into two new bands, yielding a total of four equally spaced bands. The filter used for the additional division in each band is the 16-tap QMF of Johnston designated C in Ref. 5. Figure 3 shows a block diagram for this coder.

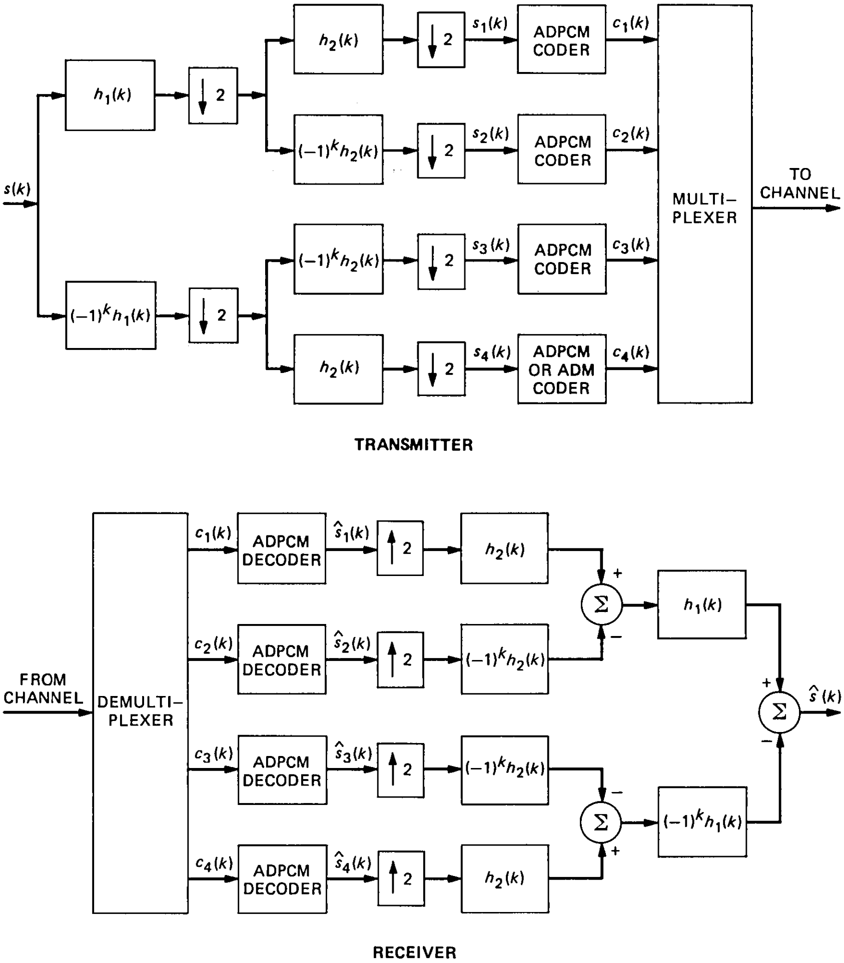


Fig. 3—Four-band sub-band coder used in simulation.

Once more the bands are quantized using ADPCM or ADM. Our measurements of average correlation for speech data showed correlations of 0.4, 0, 0, and 0.8 for the four bands going from lowest to highest in frequency. The fourth band (3000 to 4000 Hz) has actually been bandpass-filtered to cut off at 3200 Hz. As a result, it contains little power and can be ignored for low-bit rate coders. The correlations of the two middle bands are zero, reflecting that the long-term average of the speech spectrum from 1000 to 3000 Hz is flat. If a prediction coefficient of zero is used with ADPCM, the result is APCM. Thus, the two middle bands are APCM-encoded. The largest amount of power is in the first band; therefore, it receives the most bits.

The bit allocations found to be the best by the same segmental SNR measurements and casual listening process were as follows:

16 kb/s 4,2,2,0 (bands 1 to 4)
20 kb/s 5,2,2,1 (ADM on band 4)
24 kb/s 5,4,2,1
28 kb/s 7,4,2,1
32 kb/s 7,4,3,2.

The greatest amount of error occurs in the lowest band. Even at the high rates (28 and 32 kb/s) this error is still perceptible as a low rumbling noise. However, it was found that a high-pass filter with a cutoff of 200 Hz eliminated this problem. The filter used was a 121-tap FIR filter. Table I gives the coefficients, and Fig. 4 shows the frequency response. A much smaller IIR filter could also be used to do the same job.[2] Note that the above bit assignments were made without using the FIR filter. With a high-pass filter, fewer bits could be allocated to the lowest band and more to bands two and three.

2.4 Relative complexity of designs

The ADPCM designs for 16, 24, and 32 kb/s have already been implemented on the DSP.[1] The combined encoder and decoder algorithms use 48 percent of the DSP real-time capability for a sampling

Table I—Coefficients for symmetric FIR high-pass filter

−.153203E-01	−.488296E-03	−.427259E-03	−.305185E-03	−.152593E-03
.610370E-04	.305185E-03	.610370E-03	.946074E-03	.131230E-02
.170904E-02	.213630E-02	.259407E-02	.305185E-02	.350963E-02
.396740E-02	.442518E-02	.485244E-02	.524918E-02	.558488E-02
.589007E-02	.610370E-02	.625629E-02	.631733E-02	.628681E-02
.616474E-02	.592059E-02	.555437E-02	.509659E-02	.448622E-02
.37537BE-02	.283822E-02	.183111E-02	.732444E-03	−.549333E-03
−.192267E-02	−.344859E-02	−.506607E-02	−.677511E-02	−.857570E-02
−.104679E-01	−.123905E-01	−.144047E-01	−.164190E-01	−.184332E-01
−.204779E-01	−.224921E-01	−.244453E-01	−.263680E-01	−.281991E-01
−.299387E-01	−.315867E-01	−.331126E-01	−.344554E-01	−.356761E-01
−.367443E-01	−.375988E-01	−.383007E-01	−.387890E-01	−.390942E-01
.100000E-01				

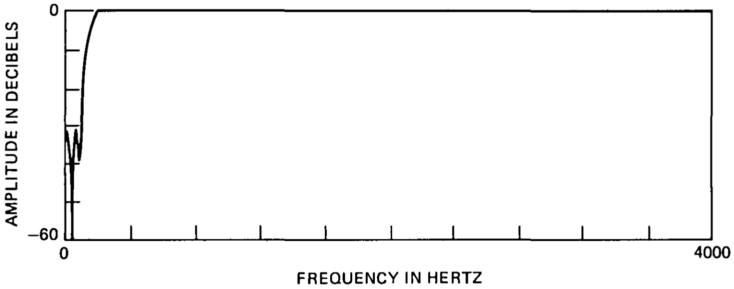


Fig. 4—Frequency response of high-pass filter for 4B-SBC.

rate of 8 kHz. An even lesser percentage of RAM and ROM memory is used. The 2B-SBC based on the design parameters reported here has also been implemented on the DSP chip.[2] It uses 98 percent of the real-time capability and 78 percent of the RAM memory. It includes an IIR bandpass filter for the input. The 4B-SBC algorithm is planned for implementation in the near future. Since all of the major portions of the 4B-SBC have been programmed already for the 2B-SBC implementation, it is possible to project how much of the DSP will be used. Both the transmitter and the receiver will require a DSP and each will use about the same fractions of real-time capability and RAM as the complete 2B-SBC algorithm. Therefore, so we might classify the three coders as having complexities of 0.5, 1, and 2, respectively.

III. RELATIVE PERFORMANCE OF THE CODERS

Since the SBC designs are more complex, a demonstration of their improved performance over ADPCM was needed to justify their implementation on the DSP. To demonstrate their relative performance all three coders were simulated on a laboratory computer. Each processed speech from a stored file. The results were evaluated by both an objective and subjective measure. The objective measure was segmental SNR, while the subjective measure was a forced-choice, subjective (A–B) comparison test in which all possible coders were compared.

Six phonetically balanced sentences were used for evaluating the coders. Three were spoken by male speakers and three by females. They were recorded using a linear microphone. They were band-limited from 200 to 3200 Hz and sampled at 8000 Hz using a 15-bit linear quantizer.

3.1 Segmental signal-to-noise ratio results

In computing segmental SNR measurements, blocks of speech of 32 ms were used. The ADPCM coder was compared with the original input speech. The SBC coders were compared with reassembled speech which had been processed by the appropriate QMF filtering, but with no quantization. These slightly modified speech signals cannot be distinguished from the original in casual listening. Without them it would be difficult to make a fair comparison of the three coders on the basis of SNR. The measurements on 4B-SBC were made before the 121-tap FIR high-pass filtering.

The results of these measurements are summarized in Fig. 5. They show that the more complex SBC coders have a definite advantage over ADPCM at the lower-bit rates. Interestingly at 32 kb/s, ADPCM beats both of the more complex coders. The 4B-SBC maintains a fairly constant 2-dB advantage over 2B-SBC. In terms of bit rate this translates to 4 kb/s. At the low rates, the 4B-SBC has about a 6-kb/s advantage over ADPCM.

3.2 Subjective testing of the three coders

An A–B comparison test was performed to rank the three coders. Each coder at each rate was compared twice against every other coder at every rate, as well as against the original. In the two comparisons of the two coders, each one was played in first position once. There were 12 participants in the test and altogether there were 240 comparisons. The test was broken down into two parts, one with 110 comparisons, the other with 130. The participants listened over headphones in a soundproof booth. The participants were also broken down into two groups of six. If one group listened to a particular A–B comparison with a female speaker the other group heard a sentence with a male speaker and vice versa. Thus, we attempted to make a totally balanced and unbiased test.

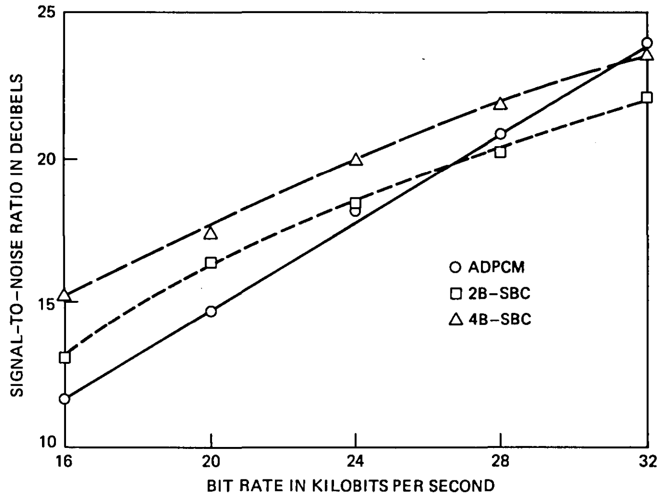


Fig. 5—Segmental SNR measurements for three coders.

Table II—Coder versus coder ratings
(Number represent percent obtained by coder listed at left of line against other coders)

		ADPCM					2B-SBC					4B-SBC				
	Orig	16	20	24	28	32	16	20	24	28	32	16	20	24	28	32
ADPCM																
16	0	—	17	0	0	0	33	8	0	0	0	0	0	0	0	0
20	0	83	—	4	0	0	79	8	4	0	0	8	0	4	0	0
24	0	100	96	—	4	21	92	63	17	13	17	75	21	21	8	13
28	8	100	100	96	—	21	96	63	67	54	33	58	63	21	25	13
32	25	100	100	79	79	—	100	87	100	63	50	100	83	58	50	54
2B-SBC																
16	0	67	21	8	4	0	—	4	0	0	0	8	0	0	0	0
20	0	92	92	37	37	13	96	—	33	8	13	37	29	13	8	13
24	4	100	96	83	33	0	100	67	—	25	37	58	50	37	29	13
28	8	100	100	87	46	37	100	92	75	—	50	71	63	33	42	29
32	17	100	100	83	67	50	100	87	63	50	—	100	96	58	46	29
4B-SBC																
16	8	100	92	25	42	0	92	63	42	29	0	—	42	13	4	4
20	4	100	100	79	37	17	100	71	50	37	4	58	—	29	13	0
24	4	100	96	79	79	42	100	87	63	67	42	87	71	—	46	25
28	21	100	100	92	75	50	100	92	71	58	54	96	87	54	—	25
32	37	100	100	87	87	46	100	87	87	71	71	96	100	75	75	—

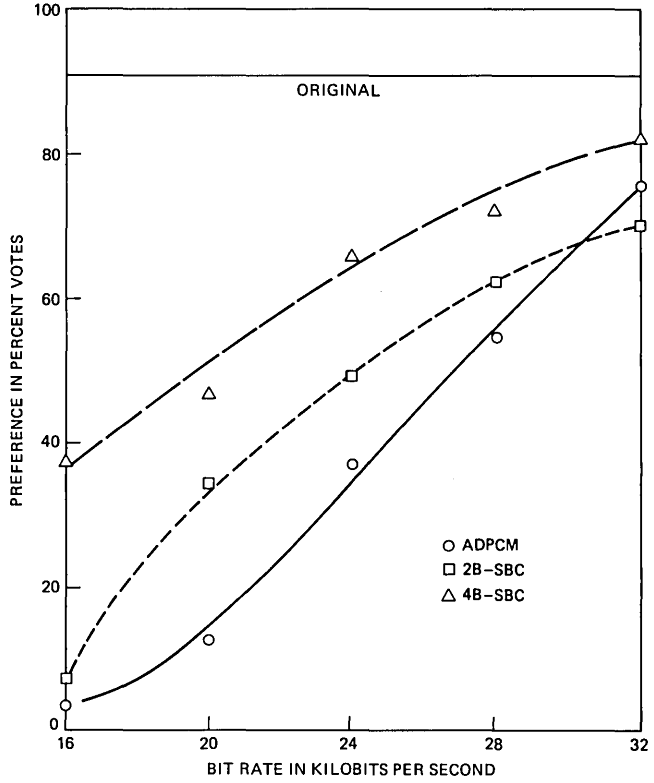


Fig. 6—Overall preference rankings for three coders.

Table II gives the individual coder versus coder comparisons. In addition, an overall preference ranking was computed based on the total number of votes received by each coder. In all, a total of 360 votes could be received by any coder. Figure 6 shows the percentage of the 360 possible votes received for each coder. This result is in good agreement with the results of Fig. 4. For example, both sub-band coders show adantage over ADPCM at the low rates and ADPCM catches up or passes them at the high rates.

Some of the more significant results are the following:

(*i*) The 4B-SBC has an 8-kb/s perceptual advantage over ADPCM at the low rates. The 24-kb/s ADPCM has been used for voice storage and playback systems.[8] This result indicates that 16-kb/s 4B-SBC could be substituted at a 33 percent savings in storage or, equivalently, a 50 percent increase in message storage capability. Moreover, at 20-kb/s, 2B-SBC has a 4-kb/s perceptual advantage over ADPCM.

(*ii*) Although 4B-SBC lost to ADPCM at 32 kb/s in SNR measurements, it beat ADPCM in the subjective tests. In addition, in direct comparisons with the original, 32-kb/s 4B-SBC received 37.5 percent of the votes, an almost equipreferential rating. This indicates it is high quality. Since 32-kb/s ADPCM is often described as toll quality, then 32-kb/s 4B-SBC also deserves this ranking.

(iii) The 2B-SBC seems to provide a good alternative to ADPCM at the low bit rates for a modest increase in complexity.

IV. CONCLUSIONS

We have presented measurement data and simulation results for use in implementing two sub-band coders on the DSP. This data has already been used to design and implement the 2B-SBC and is being used for a planned 4B-SBC implementation. Simulations of these candidate coders were made on a laboratory computer. The results of these simulations indicate 2B-SBC and 4B-SBC have important advantages over ADPCM at low bit rates. This advantage is as much as 8 kb/s for 4B-SBC and 4 kb/s for 2B-SBC. The 16-kb/s 4B-SBC could be substituted for 24-kb/s ADPCM in a voice storage and playback system. In addition, 4B-SBC is rated as better quality at 32 kb/s than at 32-kb/s ADPCM.

Since the complexity of these coders is within an order of magnitude of ADPCM they should be considered as viable alternatives.

REFERENCES

1. J. R. Boddie et al., "Digital Signal Processor: Adaptive Differential Pulse-Code-Modulation Coding," B.S.T.J., this issue, Part 2.
2. R. E. Crochiere, "Digital Signal Processor: Sub-band Coding," B.S.T.J., this issue, Part 2.
3. P. Cummiskey, N. S. Jayant, and J. L. Flanagan, "Adaptive quantization in differential PCM coding of speech," B.S.T.J., *52* (September 1973), pp. 1105–18.
4. J. D. Johnston and D. J. Goodman, "Multipurpose hardware for digital coding of audio signals," IEEE Trans. on Comm., *COM-26* (November 1978), pp. 1785–8.
5. J. D. Johnston, "A filter family designed for use in quadrature mirror filter banks," Proc. Int. Conf. Acoustics, Speech and Signal Processing, 1980 (April 1980), pp. 291–4.
6. J. D. Johnston and R. E. Crochiere, "An all digital 'commentary grade' sub-band coder," J. Audio Engineering Society, *27,* No. 11 (November 1979), pp. 855–65.
7. N. S. Jayant, "Digital coding of speech waveforms: PCM, DPCM and DM quantizers," Proc. IEEE, *62* (May 1974), pp. 611–32.
8. L. H. Rosenthal et al., "A multiline computer voice response system using ADPCM coded speech," IEEE Trans. on Acoustics, Speech and Signal Processing, *ASSP-22* (October 1974), pp. 339–52.

CONTRIBUTORS TO THIS ISSUE

Sudhir R. Ahuja, B. Tech. (Electrical Engineering), 1972, I.I.T. Bombay, India; M.S. (Electrical Engineering), 1974, Ph.D. (Electrical Engineering), 1977, Rice University. Mr. Ahuja has been working in the field of associative processing, multiprocessing, and networking. He has built an associative processor for information retrieval and has designed several high-speed buses. He is currently involved in experiments on distributed processing. Member, IEEE, ACM.

Richard V. Cox, B.S. (Electrical Engineering), 1970, Rutgers University; M.A., 1972, Ph.D., (Electrical Engineering), 1974 Princeton University; Aerospace Corporation, 1973–1977; Assistant Professor, Rutgers University, 1977-1979; Bell Laboratories, 1979—. Mr. Cox is a member of the Acoustics Research Department. His current research interests are in digital speech coding and real-time speech coding systems.

Tore E. Dalenius, Ph.D. (Statistics), 1957, University of Uppsala (Sweden); University of Stockholm, 1958-1971; Brown University, 1972—. Mr. Dalenius has been a Professor of Statistics at the University of Stockholm since 1965 and has been a Visiting Professor at Brown University since 1972. He is also a consultant to Bell Laboratories. Mr. Dalenius has contributed extensively to the field of survey sampling. He is currently the president of the International Association of Survey Statisticians.

Geng-Seng Fang, B.S.E.E., 1967, National Taiwan University; Ph.D. (E.E.), 1971, Princeton University; Computer Sciences Corporation, 1971-72; Bell Laboratories, 1972—. At Bell Laboratories, Mr. Fang has worked on high-speed digital transmission, wide-band analog transmission, protection switching, microprocessor applications, satellite systems, and echo cancelers. During 1977–79, he was with the Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan, China.

Arie Feuer, B.Sc. (Mech. Eng.), 1967, M.Sc. (Mechanical Engineering), 1973, Technion, Israel; Ph.D. (Control Systems Engineering), 1978, Yale University. 1970-1973 Bell Laboratories, 1978—. Since joining Bell Laboratories Mr. Feuer has been involved in telephone network measurement planning and implementations. He is actively involved in research in control system theory and currently looks into its possible applications for network measurements and operations.

Ytzhak L. Levendel, B.S.E.E., 1971, Technion-Israel; M.S.C.S., 1974, The Weitzman Institute of Science; Ph.D., 1976, University of Southern California; Bell Laboratories 1976—. Mr. Levendel has done research in fault diagnosis and is currently involved in the development of a logic and test design aid system. Member, IEEE, Eta Kappa Nu.

Sing-Hsiung Lin, B.S.E.E., 1963, National Taiwan University; M.S.E.E., 1966, and Ph.D., 1969, University of California, Berkeley; Bell Laboratories, 1969—. Mr. Lin is supervisor of the Digital Network Engineering Group. He developed and constructed the 11-GHz radio path engineering charts for 200 major U.S. locations which have been used widely by the Bell System operating companies and AT&T Long Lines in the engineering of new radio routes. He has also been engaged in research on digital and analog transmissions in twisted wire pair cables. Member, IEEE, Sigma Xi.

Premachandran R. Menon, B.Sc. (Electrical Engineering), 1954, Banaras Hindu University; Ph.D. (Electrical Engineering), 1962, University of Washington; Bell Laboratories, 1963—. Mr. Menon has done research in switching theory and fault diagnosis and is currently involved in the development of a logic simulation system. Member, IEEE.

Craig E. Miller, B.S.E.E., 1966, M.S.E.E., 1968, and Ph.D., 1977, Northwestern University; Bell Laboratories 1968—. Mr. Miller has worked on the characterization of digital integrated circuits and the design of bipolar ROMs for microprocessor applications. He has also worked on the characterization of dynamic RAMs for mainstore memory systems. More recently, he has developed logic gate models for the analysis of pulse-excited gate transmission line systems. He is presently responsible for timing analysis models for commercial integrated circuits and the electrical specifications for specific devices.

Cory S. Myers, B.S., M.S. (Electrical Engineering and Computer Science), 1980, Massachusetts Institute of Technology, Cambridge; Bell Laboratories, 1977—. At Bell Laboratories, Mr. Myers initially worked on computer graphics, digital circuit design, and dynamic programming for speech recognition. He is currently in the digital signal processing group, where his interests include speech processing, recognition, and digital signal processing.

Vijayan N. Nair, B. Econ. (Hons.), 1972, University of Malaya; Ph.D. (Statistics), 1978, University of California, Berkeley; Rubber Research Institute of Malaysia, 1972–1974; Bell Laboratories, 1978—. At Bell Laboratories, Mr. Nair has been working on problems in survey sampling and nonparametric and large-sample statistical inference.

Herman M. Presby, B.A., 1962, and Ph.D., 1966, Yeshiva University; Research Scientist, Columbia University, 1966–1968; Assistant Professor Physics, Belfer Graduate School of Science, Yeshiva University, 1968–1972; Bell Laboratories, 1972—. Mr. Presby is engaged in the studies on the properties of optical fiber waveguides.

Lawrence R. Rabiner, S.B. and S.M., 1964, Ph.D. (Electrical Engineering), Massachusetts Institute of Technology; Bell Laboratories, 1962—. From 1962 through 1964, Mr. Rabiner participated in the cooperative plan in electrical engineering at Bell Laboratories. He worked on digital circuitry, military communications problems, and problems in binaural hearing. Presently, he is engaged in research on speech communications and digital signal processing techniques. He is coauthor of *Theory and Application of Digital Signal Processing* (Prentice-Hall, 1975) and *Digital Processing of Speech Signals* (Prentice-Hall, 1978). Former President, IEEE, ASSP Society; former Associate Editor, G-ASSP Transactions; former member, Technical Committee on Speech Communication of the Acoustical Society. Member, G-ASSP Technical Committee of the Acoustical Society, G-ASSP Technical Committee on Speech Communication, IEEE Proceedings Editorial Board, Eta Kappa Nu, Sigma Xi, Tau Beta Pi. Fellow, Acoustical Society of America, IEEE.

Attilio J. Rainal, University of Alaska; University of Dayton, 1950–52; B.S.E.Sc., 1956, Pennsylvania State University; M.S.E.E., 1959, Drexel University; D. Eng., 1963, Johns Hopkins University; Bell Laboratories, 1964—. Mr. Rainal's early work involved research on noise theory with application to detection, estimation, radiometry, and radar theory. He has also been engaged in the analysis of FM communication systems. His more recent work includes studies of crosstalk on multilayer boards, voltage breakdown, and current-carrying capacity of printed wiring, and electromagnetic compatibility. Member, Tau Beta Pi, Eta Kappa Nu, Sigma Tau, Pi Mu Epsilon, Sigma Xi, IEEE.

Dhiraj K. Sharma, B. Tech. (Electrical Engineering), 1971, I.I.T. Kanpur, India; M.S. and Ph.D. (Electrical Engineering), 1972 and 1975, California Institute of Technology; Bell Laboratories 1975—. Mr. Sharma has worked on efficient encoding of video signals and identification of structural parameters of tall buildings. His current interests are in screen-based human interfaces, communication and synchronization in distributed systems, and programming languages. Member, ACM, Sigma Xi.

PAPERS BY BELL LABORATORIES AUTHORS

COMPUTING

CARL—Experience of an Application Using Clusters. E. Levinson, L. S. Levy, and J. B. Salisbury, AFIPS Proc Natl Computer Conf (May 4, 1981), pp 241–8.

Communication System Architecture. R. F. Rosin, Amer Fed Inform Processing Societies Office Automation Conf Digest (March 1981), pp 269–70.

Data Communication Using the Telecommunication Network. M. F. Slana and H. R. Lehman, Computer, Magazine of the IEEE Computer Society, *14*, No. 5 (May 1981), pp 73–88.

Hierarchical Schemata for Relational Databases. Y. E. Lien, Assn Computing Machinery Trans on Database Systems, *6*, No. 1 (March 1981), pp 48–69.

An Integrated Multiprocessing Array for Time Warp Pattern Matching. B. Ackland, N. Weste, and D. J. Burr, Proc 8th Annual Symp on Computer Architecture (May 1981), pp 197–215.

A Note on the Leech Lattice as a Code for the Gaussian Channel. N. J. A. Sloane, Information and Control, *46* (1980), pp 270–2.

Office of the Future—Several Viewpoints. C. Stockbridge, I. F. Chang, E. C. Greanias, G. G. Pick, and B. A. Saltzer, Proc Soc Info Display, *22*, No. 1 (1981), pp 19–22.

A Software Study Using Halstead Metrics. C. T. Bailey and W. L. Dingee, Performance Eval Rev, Proc 1981 ACM/Sigmetrics Conf, *10*, No. 1 (March 1981), pp 189–97.

A Stored Program Controlled Triport^{TM} UPS. H. E. Menkes, Proc 3rd Int Telecommun Energy Conf, IEE Conf Pub No. 196 (1981), pp 210–5.

Teleconferencing, In and Outside Office Automation. C. Stockbridge, Amer Fed Inform Processing Societies Office Automation Conf Digest (March 1981), pp 267.

Telematics: The Integration of Computing and Communications. D. Gillette, Computerworld Extra, *XV*, No. 11a (March 1981), pp 33–4.

ENGINEERING

Digital Satellite Communications Problems and Possibilities. V. I. Johannes, Proc IEEE Conf Communications Techniques Seminar—Digital Communications (March 24, 1981).

The DSP—Architecture and Applications Overview. I. I. Eldumiati, R. N. Gadenz, and N. Sachs, Proc 1981 IEEE Int Symp Circuits and Systems (April 1981), pp 888–90.

The Fault-Tolerant 3B-20 Processor. L. E. Gallaher and W. N. Toy, Amer Fed Inform Processing Societies Proc Natl Computer Conf (May 4, 1981), pp 41–8.

Glassy Metals. H. S. Chen, Report on Progress in Physics, *43* (April 1980), pp 353–432.

Hard-Limited Versus Linear Combining for Frequency-Hopping Multiple-Access Systems in a Rayleigh Fading Environment. O. Yue, IEEE Trans Vehicular Technology, *VT-30*, No. 1 (February 1981), pp 10–4.

Index-of-Refraction Matching Materials for Optical Fiber Splicing. S. C. Mettler and M. R. Gotthardt, Proc 3rd Int Conf Integrated Optics and Optical Fiber Commun (April 1981), pp 94–6.

Interpolation and Decimation of Digital Signals—a Tutorial Review. R. E. Crochiere and L. R. Rabiner, Proc of the IEEE, *69*, No. 3 (March 1981), pp 300–31.

Measurements of Second Sound in Partially Spin-Polarized ^{3}He-^{4}He Solutions. D. S. Graywoll and M. A. Paalanen, Phys Rev Lett, *46*, No. 19 (May 11, 1981), pp 1292–5.

Neutron Activation Analysis Study of the Sources of Transition Group Metal Contamination in the Silicon Device Manufacturing Process. P. F. Schmidt and C. W. Pearce, J Electrochem Soc, *128*, No. 3 (March 1981), pp 630–7.

Neutron Activation Study of a Gettering Treatment for Czochralski Silicon Substrates. L. E. Katz, P. F. Schmidt, and C. W. Pearce, J Electrochem Soc, *128,* No. 3 (March 1981), pp 620–4.

Optical Fiber Components Using Grin-Rod Lenses. R. H. Knerr, Topical Meeting on Gradient Index Optical Imaging Systems, Optical Soc of Amer Tech Digest (May 4–5, 1981), Paper No. TM C2-1.

Performance of Transform and Sub-band Coding Systems Combined with Harmonic Scaling of Speech. D. Malah and R. E. Crochiere, IEEE Trans Acoustics, Speech, and Signal Processing, *ASSP-29* (April 1981), pp 273–83.

Radio System Interference From Geostationary Satellites. P. E. Butzien, IEEE Trans Commun, *COM-29,* No. 1 (January 19, 1981), pp 33–40.

Reliability Evaluation of Aluminum-Metallized MOS Dynamic RAMs in Plastic Packages in High Humidity and Temperature Environments. K. M. Striny and A. W. Schelling, Proc 1981 31st Electronic Components Conf.

SCARAB (Submersible Craft Assisting Recovery and Burial). H. R. Lunde, G. A. Reinold, and P. A. Yeisley, Jr., Proc 1981 Offshore Technology Conf (May 1981).

A Single-Chip CMOS PCM Codec with Filters. B. K. Ahuja, M. R. Dwarakanath, T. E. Seidel, and D. G. Marsh, Int Solid State Circuits Conf 1981, Digest of Technical Papers (February 1981), pp 242–3.

Strength Characterization of Multikilometer Silica Glasses. F. V. DiMarcello, D. L. Brownlow, and D. S. Shenk, 3rd Int Conf Integrated Optics and Optical Fiber Commun Tech Digest (April 1981), pp 26–7.

Use of Intracavity Filters for Optimization of Far-Infrared Free Electron Lasers. E. D. Shaw and C. K. N. Patel, Phys Rev Lett, *46,* No. 5 (February 2, 1981), pp 332–5.

MANAGEMENT and ECONOMICS

Existence of Sustainable Prices for Natural Monopoly Outputs. W. Sharkey, Bell J Econ, *12,* No. 1 (Spring 1981), pp 144–54.

Field Evaluating Documents: What Can We Really Expect? E. S. Brendel, Proc 28th Int Tech Commun Conf (May 1981), pp E7–E10.

Selection of MIL-STD-105D Plans Based on Costs. B. S. Liebesman, Trans 35th Annual Amer Soc Quality Control Quality Congress and Exposition (May 27–29, 1981), pp 475–84.

The Universal Sampling Plan. B. Hoadley, Trans 35th Annual Amer Soc Quality Control Quality Congress (May 27-29, 1981), pp 80–7.

MATHEMATICS

Aleph-Projective Spaces. C. W. Neville and S. P. Lloyd, Ill J Math, *25* (1981), pp 159–68.

PHYSICAL SCIENCES

Amorphous SiO_2-Co Spin Glasses. J. J. Hauser, Solid State Commun, *37* (January 1981), pp 344–51.

Carbonyl Sulfide: Potential Agent of Atmospheric Sulfur Corrosion. T. E. Graedel, G. W. Kammlott, and J. P. Franey, Science, *212,* No. 4495 (May 8, 1981), pp 663–5.

A Complete CO Map of a Spiral Arm Region in M31. F. Boulanger, A. A. Stark, and F. Combes, Astronomy and Astrophysics, *93* (January 1, 1981), pp L1–L4.

Contrast Enhancement in Multicomponent Polymer Systems. M. J. Bowden, J Appl Poly Sci, *26* (1981), pp 1421-6.

Crystallization of the β-Phase of Poly(Vinylidene Fluoride) From the Melt. A. J. Lovinger, Polymer, *22*, No. 3 (March 1981), pp 412–3.

The Effects of Phosphorus Diffusion Cooling Rate on I^{2}L Gain. B. L. Morris, Solid State Elec, *23* (June 1980), pp 457–65.

Electron Microscopy and Failure Analysis. R. B. Marcus and T. T. Sheng, Proc 19th Annual Reliability Phys Symp (April 1981).

Fundamental Aspects of Czochralski Silicon Crystal Growth for VLSI. K. E. Benson, W. Lin, and E. P. Martin, 4th Intl Symp on Silicon Materials and Technology, *Semiconductor Silicon 1981*, edited by H. R. Huff and R. J. Kriegler, Princeton: Electrochem Soc, May 1981, pp 33–48.

The Galactic Rotation Curve to R 18 Kpc. L. Blitz, M. Fich, and A. A. Stark, *Interstellar Molecules*, Proc Int Astronomical Union Symp, No. 87, edited by B. H. Andrew, Boston: D. Reidel Publishing, 1980, pp 213–20.

Isotopic Species of HCO$^+$ in Giant Molecular Clouds. A. A. Stark, Astrophysical J, *245*, No. 1 (April 1, 1981), pp 99–104.

Long-Lived High-Rydberg Molecules Formed by Electron Impact: H_2, D_2, N_2, and CO. S. M. Tarr, J. A. Schiavone, and R. S. Freund, J Chem Phys, *74* (March 1, 1981), pp 2869–78.

Metal Chalcogenides as Reversible Cathodes in Lithium Cells and Their Future in Telecommunications. J. Broadhead, F. A. Trumbore, and S. Basu, J Electroanal Chem, *118* (1981), pp 241–9.

Molecular Mechanism for $\alpha \rightarrow \delta$ Transformation in Electrically Poled Poly(Vinylidene Fluoride). A. J. Lovinger, Macromolecules, *14*, No. 1 (February 1981), pp 225–7.

Ohmic Contacts on n-GaAs Produced by Spark Alloying. R. D'Angelo and P. A. Verlangieri, Elec Lett, *17*, No. 8 (April 16, 1981), pp 290–1.

Optical Studies of Structural Phase Transitions. P. A. Fleury and K. B. Lyons, *Structural Phase Transitions*, Topics in Current Physics, edited by K. Muller and H. Thomas, Berlin: Sprenger-Verlag, 1980, pp 9–92.

Perturbation of Membrane Structure by Uranyl Acetate Labeling. V. A. Parsegian, R. P. Rand, and J. Stamatoff, Biophysical J, *33*, No. 3 (March 1981), pp 475–8.

Phase Transitions, Critical Phenomena and Instabilities. P. A. Fleury, Science, *211* (1981), pp 125–31.

The Physics and Chemistry of the Lithographic Process. M. Bowden, J Electrochem Soc *128*, No. 5 (1981), pp 195C–214C.

Plasma Etching of Silicon and Silicon Dioxide With Hydrogen Fluoride Mixtures. G. Smolinsky, R. P. H. Chane, and T. M. Mayer, *Plasma Processing*, edited by R. E. Frieser and C. J. Mogab, Proc Electrochem Soc, *81-1*, 1981, pp 120–4.

Spin-Glass-Ferromagnetic Multicritical Point in Amorphous Fe-M$_n$ Alloys. M. B. Salamon, K. V. Rao, and H. S. Chen, Phys Rev Lett, *44* (March 1980), pp 596–9.

Structural Phase Transitions: Defects and Dynamics. P. A. Fleury, Proc 3rd Int Topical Conf on Lattice Defects in Ionic Crystals, J. de Physique, *41* (July 1980), pp C6-419.

The Surface Photovoltage of Squarylium Dye Films. M. E. Musser and S. C. Dahlberg, Applic Surf Sci, *5* (1980), pp 28–36.

Theory of the Cholesteric Blue Phase. S. Meiboom, J. P. Sethna, P. W. Anderson, and W. F. Brinkman, Phys Rev Lett, *46*, No. 8 (May 4, 1981), pp 1216–9.

Thermal Conductivity Measurements in Liquid ^{4}He Below 0.7K. D. S. Greywall, Phys Rev, *23*, No. 5 (March 1, 1981), pp 2152–68.

Underwater Sound From Surface Waves According to the Lighthill-Ribner Theory. S. P. Lloyd, J Acoust Soc of Amer, *69* (1981), pp 425–35.

Visible Absorption Spectrum of Liquid Ethylene. E. T. Nelson and C. K. N. Patel, Proc Natl Acad Sci, *78*, No. 2 (February 1981), pp 702–5

CONTENTS, OCTOBER 1981

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*, *Current Contents/ Engineering, Technology & Applied Sciences*, *Current Index to Statistics*, *Current Papers in Electrical & Electronic Engineering*, *Current Papers on Computers & Control*, *Electronics & Communications Abstracts Journal*, *The Engineering Index*, *International Aerospace Abstracts*, *Journal of Current Laser Abstracts*, *Language and Language Behavior Abstracts*, *Mathematical Reviews*, *Science Abstracts* (*Series A, Physics Abstracts*; *Series B, Electrical and Electronic Abstracts*; *and Series C, Computer & Control Abstracts*), *Science Citation Index*, *Sociological Abstracts*, *Social Welfare, Social Planning and Social Development*, 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