On Talker Verification Via Orthogonal Parameters

ROBERT E. BOGNER, MEMBER, IEEE

Abstract—The method of orthogonal parameters uses the eigenvalues and eigenvectors of the matrix of covariances of measurements made on a population of utterances from a given talker to transform measurements made on an utterance from an unknown talker. The values include a subset of the principle components (obtained by Hotelling transformation), plus other values dependent on the variances of the principle components. Variations of the method are given.

Earlier work applied the method to measurements which were sets of linear prediction coefficients, reflection coefficients, and logarithmic areas derived from the speech waveform, and suggested that highly accurate speaker verification could be achieved, independent of the speech transmission system, and of the text. Present work shows that there is serious dependence on the transmission system, and that accuracies in the range 90–95 percent are typical with utterances of about 2 s duration.

Investigation of a new distance measure based on covariances, combined with study of the factors influencing distortion in telephone transmission, resulted in a new method based on covariances of logarithmic spectral estimates. The results of this method, giving accuracies of about 95 percent, were robust in the presence of transmission coloration. Some remaining research problems are defined.

I. Introduction

SOME earlier experiments [1] on speaker verification by observations of statistical averages of the behavior of linear prediction coefficients and related variables over a whole utterance showed promise of useful accuracy and protection from the effects of the telephone transmission medium. These experiments used a method called "orthogonal linear prediction" (OLP), based on the extraction of "orthogonal parameters." Other closely related processes used in signal processing and pattern recognition are the method of principal components [2], the Hotelling transformation [3], the Karhunen-Loeve expansion [2], and the Mahalanabis distance [4].

The orthogonal parameters, ϕ_i, $i = 1, \cdots, M$ are a set of values obtained by linear transformation of the observations x_i, $i = 1, \cdots, M$:

$$\phi_i = \sum_{j=1}^{M} b_{ji} x_j \tag{1}$$

or $$\boldsymbol{\phi} = \boldsymbol{b}^T \boldsymbol{x} \tag{2}$$

where b_{ji} is the jth element of the ith normalized eigenvector $\boldsymbol{b}_i$ of the covariance matrix $\boldsymbol{R}_x$ of the population of the x_j, and

$$R_x = [\langle (x_i - \bar{x}_i)(x_j - \bar{x}_j) \rangle] \tag{3}$$

where $\langle\ \rangle$ and the overbar indicate population averages. The covariance matrix of the ϕ_i obtained in this way, i.e.,

$$R_\phi = [\langle (\phi_i - \bar{\phi}_i)(\phi_j - \bar{\phi}_j) \rangle] \tag{4}$$

has the form

$$R_\phi = \begin{bmatrix} \lambda_1 & 0 & \cdots \\ 0 & \lambda_2 & \cdots \\ 0 & 0 & \cdots \\ & & \cdots \\ & & \cdots \lambda_M \end{bmatrix} \tag{5}$$

where the λ_i, $i = 1, \cdots, M$ are the eigenvalues of $\boldsymbol{R}_x$, which are usually ordered so that $\lambda_i > \lambda_{i+1}$. The off-diagonal terms being zero corresponds to the orthogonality of the ϕ_i, i.e., the average product in (4) being zero for $i \neq j$. The λ_i are seen to be the variances of the orthogonal parameters.

The ϕ_i, $i = 1, \cdots, M$ are sufficient to reconstruct the x_i, by inversion of (1), and since the $\boldsymbol{b}_j$ are orthonormal, then

$$x_j = \sum_{i=1}^{M} b_{ji} \phi_i \tag{6}$$

or

$$\boldsymbol{x} = \boldsymbol{b}\boldsymbol{\phi}.$$

If some of the ϕ_i as given by (1) have negligibly small variations from their means $\bar{\phi}_i$ as evidenced by small λ_i, then the partial sum

$$\hat{x}_j = \sum_{i=1}^{M_2} b_{ji} \phi_i + \sum_{i=M_2+1}^{M} b_{ji} \bar{\phi}_i$$

where $M_2 < M$ may be a good approximation to x_j. The second sum is a constant vector. The orthogonal property ensures that this approximation has the least mean-squared error for a given M.

A particular attraction of the orthogonality of the parameters arises in consideration of the M-dimensional probability density function $p_x(\boldsymbol{x})$:

$$p_x(\boldsymbol{x}) = (2\pi)^{-M/2} |R_x|^{-1} \exp [-(\boldsymbol{x} - \bar{\boldsymbol{x}})^T R_x^{-1} (\boldsymbol{x} - \bar{\boldsymbol{x}})/2] \tag{7}$$

$$= (2\pi)^{-M/2} |R_\phi|^{-1} \exp [-(\boldsymbol{\phi} - \bar{\boldsymbol{\phi}})^T R_\phi^{-1} (\boldsymbol{\phi} - \bar{\boldsymbol{\phi}})/2] \tag{8}$$

in which the calculation of the exponent is simplified by the form of R_ϕ, (5). The exponent is then $D_M/2$ with

Manuscript received September 28, 1979; revised May 5, 1980 and September 6, 1980.

The author was with the Acoustics Research Department, Bell Laboratories, Murray Hill, NJ 07974. He is now with the Department of Electrical Engineering, University of Adelaide, Adelaide, Australia.

0096-3518/81/0200-0001$00.75

$$D_M = \sum_{i=1}^{M} \frac{(\phi_i - \bar{\phi}_i)^2}{\lambda_i}, \tag{9}$$

a Euclidian (distance)^2 since the ϕ_i are orthogonal. We call this D_M because it is often known as a Mahalanabis distance in either form, (8) or (9) [4].

In practical talker verification tasks values of y_i, $i = 1, \cdots, M$ are observed, corresponding to the x_i for an unknown utterance and we test the hypothesis that the utterance is from a proposed talker characterized by $\bar{x}$ and R_x, by use of (7) or the more convenient form (8). Thus we calculate

$$D_M = \sum_{i=1}^{M} \frac{(\phi_{iy} - \bar{\phi}_i)^2}{\lambda_i} \tag{10}$$

where ϕ_{iy} is the transformed version of the y_i obtained from

$$\boldsymbol{\phi}_y = \boldsymbol{b}^T y \tag{11}$$

using the $\boldsymbol{b}$ for the proposed talker. The y used in such a test is the best estimate available for the utterance under test, and is in the experiments reported here, the mean of over a hundred observations. The $\bar{x}$ or $\bar{\boldsymbol{\phi}}$ and $\boldsymbol{b}$ supposed to be characteristic of the proposed talker are estimates based on several utterances of that talker, typically 5 or 10, each utterance contributing about 150–200 observation vectors x.

In the previous investigation [1] three different types of observations were made:

1) 12 linear prediction coefficients,
2) 12 parcor (or reflection) coefficients, and
3) 12 logarithmic area coefficients.

These were calculated for the sentence "we were away a year ago" at 20 ms intervals. The sentence was spoken by 21 talkers. Other sentences were used for tests of text independence. Several transmission conditions were used, including bandlimited clean speech and speech transmitted via various local lines.

Sambur proposed from his initial experiments [1] that "the first few most significant orthogonal parameters would be indicative of the linguistic content of the utterance . . . , the least significant parameter would be indicative of the telephone media, and the remaining parameters would be indicative of the speaker . . . ." He increased the order of the model, i.e., the number of coefficients extracted, to 14 so that the variability of the media could be accommodated by the four least significant orthogonal parameters. Then he found that the identification accuracy was 94.4 percent for the metric D_M using a sum over orthogonal contributions from ϕ_i, $i = 5, \cdots, 10$ (apparently, this summing range was not stated), and 96.03 percent when the metric D_M was augmented by a contribution based on comparison of variances.

The augmentation was based on the comparison of the variances λ_i of the orthogonal parameters (5), of the reference set with the corresponding diagonal elements of the matrix

$$R_{\phi_y} = \boldsymbol{b}^T \langle (y - \bar{y})(y - \bar{y})^T \rangle \boldsymbol{b} \tag{12}$$

which would equal R_ϕ if

$$R_x = R_y = \langle (y - \bar{y})(y - \bar{y})^T \rangle,$$

i.e., if the y were exhaustive observations from the proposed population of x. Specifically, this augmentating distance D_A was given by

$$D_A = \frac{1}{2} \bar{J}_m \sum_{i=N_1}^{N_2} \left(\frac{V_i - \lambda_{im}}{\lambda_{im}} \right)^2 \tag{13}$$

where V_i is the i'th diagonal term of R_{ϕ_y} as given in (12) and $\bar{J}_m$ is the average number of frames, i.e., observation vectors, in the reference set. $2\lambda_{im}/\bar{J}_m$ is the variance in the estimation of λ_{im} and thus, the $\bar{J}_m/2$ is a form of normalization. The subscript m referred to the population of talker m. The summing range, N_1 to N_2 corresponded to that used for D_M.

Repetition of some of the experiments reported in [1] by a different laboratory [5] yielded somewhat different results. In retrospect, it is possible to account for the discrepancies by reference to inhomogeneities in the population of talkers (as discussed in Section III-C).

Initial objectives of the present work were

1) to repeat key aspects of the previous experiments,
2) to test the applicability of the method to a wider range of telephone transmission conditions, and
3) to explore the effects of one medium perturbation at a time and so to gain a better understanding of how the medium influences the parameters.

The method of OLP used in [1] will be designated as OLPA in the present paper. Study of the details of OLPA revealed that the method differed slightly from the theoretical intention described in the Introduction and [1], and a correct method OLPB was implemented, as well, to ensure that a fair representation of the performance of OLP would be obtained. OLPB includes also, as an optional addition, a distance component based on covariances [6]. This distance component was found to provide robust discrimination *when used with appropriately designed observation vectors.*

A new method, filtered logarithmic spectrum (FLS), using only variances and covariances, is described in Section VII-A in connection with the effects of variations in transmission. FLS was invented following consideration of the effects of telephone transmission on a speech signal. The direct effects of spectral distortions were proposed to be eliminatable by use of only the variances and covariances of spectral data.

II. General Principles of Calculations

This section gives the basic principles of the calculations involved. In Appendix I the differences between the calculations for OLPA and OLPB are outlined. It has been found, however, that although the experimental results tend to favor the corrected OLPB, advantage of the latter is not practically significant unless the covariance D_{OFF} is incorporated.

A. Estimation of Λ and R_x

The eigenvalues λ_i and eigenvectors $\boldsymbol{b}$ for each reference have to be estimated from utterances of the relevant talker. This involves estimating the corresponding covariance matrix.

$$R_x = [C_{i,j}]$$

with

$$C_{i,j} = \langle (x_i - \bar{x}_i)(x_j - \bar{x}_j) \rangle$$

$$= \frac{1}{N} \sum_{n=1}^{N} (x_{in} - \bar{x}_{in})(x_{jn} - \bar{x}_{jn}) \quad (14)$$

where the summation is over all the observations to be included and these relate to several utterances of the same talker, five having been used in OLPA [1]. The sum may be made up as

$$C_{i,j} = \frac{1}{N} \sum_{l=1}^{L} \sum_{n=1}^{N_l} (x_{iln} - \bar{x}_{iln})(x_{jln} - \bar{x}_{jln}) \quad (15)$$

where

$$N = \sum_{l=1}^{L} N_l,$$

L is the number of utterances, N_l is the number of observations on utterance l, and x_{inl} is the value of coefficient i in observation n for the utterance l.

B. Calculation of D_V: Use of Observed Correlations

The augmenting distance D_A used in [1], as given by (13), uses some of the information in the covariance matrix R_y. Only M values of the diagonal elements are used in D_A whereas there are $M(M+1)/2$ different values in R_y.

In [6] the theory and a procedure are developed for using all the elements of R_y. It is shown there that the appropriate metric D_C based on the correlations is

$$D_C = \frac{N}{2} \sum_{i=1}^{M} \sum_{j=1}^{M} (\Sigma_{ij} - \delta_{i,j})^2 \quad (16)$$

where $\delta_{i,j} = 1$ for $i = j$; $= 0$ otherwise and Σ_{ij} is the i,j term of

$$\Sigma = \Lambda^{-1} E^T R_y E, \quad (17)$$

Λ being the eigenvalue matrix. N is the total number of frames or observation vectors in the test utterance and is in contrast to $\bar{J}_m$.

To facilitate an understanding of the relation between the variance and covariance elements of Σ, D_C has been split into two parts:

$$D_C = D_V + D_{\mathrm{OFF}} \quad (18)$$

where

$$D_V = \frac{N}{2} \sum_{i=1}^{M} (\Sigma_{ii} - 1)^2, \quad (19)$$

comprising the diagonal terms related to variances, and

$$D_{\mathrm{OFF}} = N \sum_{i=1}^{M} \sum_{j=i+1}^{M} (\Sigma_{i,j})^2, \quad (20)$$

comprising the *off*-diagonal elements.

C. Format of Results: Combining D_M, D_V, D_{OFF}

In a verification trial the information in a reference file characteristic of talker m is compared with information extracted from the utterance of an unknown talker claiming to be m. This comparison results in a distance D, usually made up as a combination of D_M, D_V, and D_{OFF}:

$$D = K_M D_M + K_V D_V + K_{\mathrm{OFF}} D_{\mathrm{OFF}} \quad (21)$$

where the K_M, K_V, and K_{OFF} are chosen so that D is an optimal discriminating function in some sense.

For example, a threshold D_T is often chosen and the decision is made according to the rule:

Condition	Decision
$D \leqslant D_T$	unknown talker is same as reference talker
$D > D_T$	unknown talker is different from reference talker.

In the present work D_T has been set so that the probabilities of errors in making these decisions are equal and K_M, K_V, and K_{OFF} have been chosen to minimize these probabilities. More details are given below.

In the experiments it is known whether an utterance being compared with a reference is from the same or a different talker and the resultant differences are termed "same" or "different," respectively. The corresponding cumulative distributions of same- and different-distances are plotted for one set of OLPB comparisons in Fig. 1. The distribution of same-distances is cumulative from the right, while that of different-distances is cumulative from the left. Thus, the ordinate of the same-distances distribution is the probability that the same-distance is greater than or equal to the abscissa, while for the different distances the ordinate is the probability that the different-distance is less than or equal to the abscissa.

The point at which the distributions cross corresponds to the threshold distance D_T for which the probability of incorrect classification is the same for same-distances and different-distances. This point has been used to define the error rates for the experiments reported here. In each experiment it was found algorithmically from the observed distributions. The distributions were obtained by classifying the distances D_M, D_V, D_{OFF}, and D into equal intervals, there being optionally up to 200 intervals for each. The intervals between which the crossover occurred were found and the corresponding four values of error probability were used to interpolate the crossover point. Details are given in Appendix II. Thus, the threshold D_T used in each test was determined specifically to suit the process and adjustments involved.

The determination of K_M, K_V, and K_{OFF} in (21) has an important bearing on the overall performance. These coefficients were adjusted empirically after files containing D_M, D_V, and D_{OFF} had been generated, this generation being the computation-intensive part of the process. It was found that although the K's are important, the optimal values are not highly sensitive, and ±50 percent departures usually caused negligible effect. Each comparison experiment reported incorporates the results of this manual adjustment.

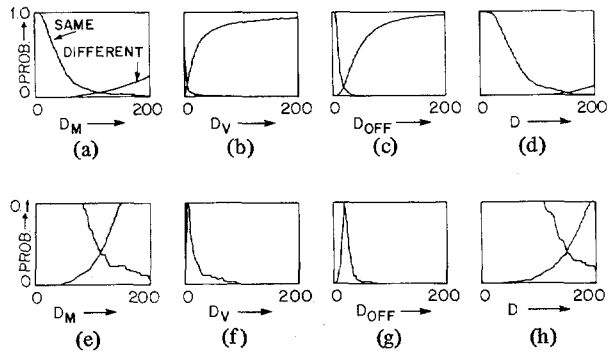

Fig. 1. Distributions of distances for OLPB for clean speech with summing range $N_1 = 1$, $N_2 = 14$. (a) D_M. (b) D_V. (c) D_{OFF}. (d) D. (e), (f), (g), (h): same as (a), (b), (c), (d) but with ordinate expanded by factor of 10.

III. Speech Databases and Experimental Designs

All utterances used were of the sentence "we were away a year ago," spoken by male talkers, and sampled at 10 000/s. All were edited to remove any sections of near silence. Two main sets of utterances were used, namely, "clean speech," recorded under studio conditions, and "telephone speech," recorded over a variety of telephone connections.

A. Clean Speech

The database was the same as that used in [1] except that only ten of the talkers were used (CS, DB, DM, JK, JM, JP, LR, MG, PB, and RS), with 30 utterances of each.

The utterances were spoken over a period of about six weeks, each subject recording one utterance each morning and afternoon, along with other sentences not used in this experiment. The first ten utterances of each talker were ignored to avoid possible familiarization effects.

The recordings were made in a quiet booth with a high quality microphone in a telephone handset. The signals were bandpass filtered to the range 100–4000 Hz before sampling and conversion at 10 000 samples/s.

B. Telephone Speech

Fifty recordings were made by each of 10 male talkers ($M0$, $M1, \cdots, M9$) at a rate of one or two per day over a period of about 6 weeks. In addition, 40 other talkers spoke one utterance each of the same sentence, to produce a set of impostor utterances, although there was no deliberate attempt to mimic the main talkers.

The sending end used a 500-type telephone in a room whose properties were typical of a normal office and the connections were made over lines dialed up via the local exchange. The received signal used was the voltage across the receiver terminals of another telephone. The signal was band limited to 100–4000 Hz and sampled at 10 000 samples/s.

Thus, the main physical sources of variability in these recordings were room noise, time variation of the carbon microphone including its position in relation to the talker's lips, and the speech paths in the exchange. Local line length, junctions, and carrier systems were not included as variables.

C. Experimental Designs

From the set of 30 utterances of each talker, two reference files were created, each based on 5 (or 10) utterances. Then the distances D_M, D_V, and D_{OFF} were calculated, as relevant, between each of the 20 references and each of the 30 utterances of every talker, excepting those utterances which had been used in generating the references. This exception meant that 5 (or 10) utterances out of each 30 were not eligible for comparison with the corresponding reference.

The distances between the reference of talker m and the utterances of talker n, $m \neq n$ are called "different" distances, and those between the reference of talker m and the utterances of talker m are called "same" distances. Thus, the numbers of distances of each class were

Same-distances: 20 references × 25 utterances = 500
or: 20 references × 20 utterances = 400
Different-distances: 20 references × 30 utterances × 9 talkers = 5400.

To appreciate the sampling implications of these designs we note that the error rates P_e are of the order of 0.07. Then if the populations were homogeneous, we can find the numbers of errors N_e and their standard deviations σ_e from Poisson statistics as in Table I.

In Table I the column headed $P_e\sigma_e/N_e$ gives the standard deviation of the expected value of the error rate for estimates based on homogeneous populations of the sizes used. For example, if the true error rate for same-distances were 0.07 then in an experiment involving the observation of 500 same-distances we expect 35 ± 6 errors, corresponding to an estimated error probability of 0.07 ± 0.012. Such calculations give an indication of the reliability that might be expected from experiments of moderate size, and show that it is not reasonable to expect close agreement with the results of brief experiments made by other workers with different sets of data.

A further complicating effect is that the talker populations need not be homogeneous. For example, when a typical set of results with clean speech was analyzed in detail it was found that about 180 of the 400-odd same-distance errors were contributed by comparisons with the references from one talker (LR, Table II, row 1). When an entirely different verification process was used (FLS, Table II, row 2) the same phenomenon was found with the same talker. One possible "explanation" is that this talker's speech is exceptionally similar to the average of the population of talkers.

Let us denote talkers of this type as "type-X." Assuming that a proportion 0.1 of the talker population is of type-X, we find that the expected number of such talkers in a sample of 10 is 1, with a standard deviation of 1. Thus, it would not be surprising to find some sets of 10 talkers with no type-X talkers, and some with 2 or 3, causing the resultant error rates to differ greatly, by factors of more than 3. This phenomenon can easily account for the widely different results reported by different authors. To reduce the effect to the order of 0.1 of the observed error rate would require numbers of test talkers approaching 1000; then, on the assumption of Poisson statis-

TABLE I
EXPECTED NUMBERS OF ERRORS N_e AND STANDARD DEVIATIONS

	N_e	σ_e	$P_e\sigma_e/N_e$
Same distances	35	$\sqrt{35} \approx 6$	0.012
Different distances	378	19	0.0035

tics the expected number of type-X talkers would be 100, with a standard deviation of 10, i.e., 0.1 of 100, and the standard deviation of the error rate would be

(10 type-X talkers) × (150 errors per type-X) = 1500

in a total expected error rate of

(1000 talkers) × (15 errors per talker) = 15 000.

Of course, the foregoing analysis is oversimplified, but it does serve to illustrate the hazards of reliance on population homogeneity when a large proportion of errors is associated with a few talkers.

However, these phenomena do not preclude the making of valuable comparative and diagnostic tests of various systems with small numbers of talkers.

IV. SPEECH PREPROCESSING

For testing of the OLP processes, vectors of 14 reflection coefficients were produced, using 300-sample Hamming windows at 100-sample intervals. The previous work [1] had shown that reflection coefficients were as good as any other LPC set.

It is common in extraction of LPC's and related coefficients to preemphasize the spectrum before analysis by filtering with a +6 dB/octave slope. In a fundamental sense, this process gives a compensation for the net -6 dB/octave slope contributed by the combination of approximately -12 dB/octave due to the glottal excitation and +6 dB/octave due to the radiation characteristic of the lips. In practical terms, it may be considered to provide a better-conditioned spectrum so that the higher frequency poles are less likely to be masked by artifacts of the necessary windowing. To test whether the effects of the 6 dB/octave equalization are significant a verification experiment was performed, using the standard set of 10 talkers, 20 references, and 30 utterances per talker, with and without preemphasis. The speech was telephone speech, band limited nominally to 300-3000 Hz (Fig. 2).

The results (Fig. 3) show that there is practically no significant difference, although there is a fairly consistent small difference in favor of the preemphasized speech for the comparisons using summing ranges 1-14 and 5-14. The advantage appears to be reversed for summing ranges 1-10 and 5-10, i.e., excluding orthogonal parameters 11-14.

Later experiments were based on preemphasized speech as there was no particular reason to do otherwise.

V. PERFORMANCE OF OLP PROCESSES

In view of the impressive results reported in [1], it was of interest to reproduce the original orthogonal linear prediction method (called OLPA), as closely as possible, despite the differences, noted in Section II and Appendix I, between it and the apparently more correct approach OLPB. Comparisons were made between the verification accuracies of OLPA and OLPB, using both the clean speech and the telephone speech databases.

TABLE II
ERRORS ASSOCIATED WITH INDIVIDUAL REFERENCES (THE NUMBERS ARE APPROXIMATE, BEING BASED ON VISUAL OBSERVATIONS FROM GRAPHS)

Talker	*CS*	*DB*	*DM*	*JK*	*JM*	*JP*	*LR*	*MG*	*PB*	*RS*
Process OLPB	24	6	60	24	30	15	180	36	18	9
Process FLS	18	0	60	4	36	36	150	0	21	9

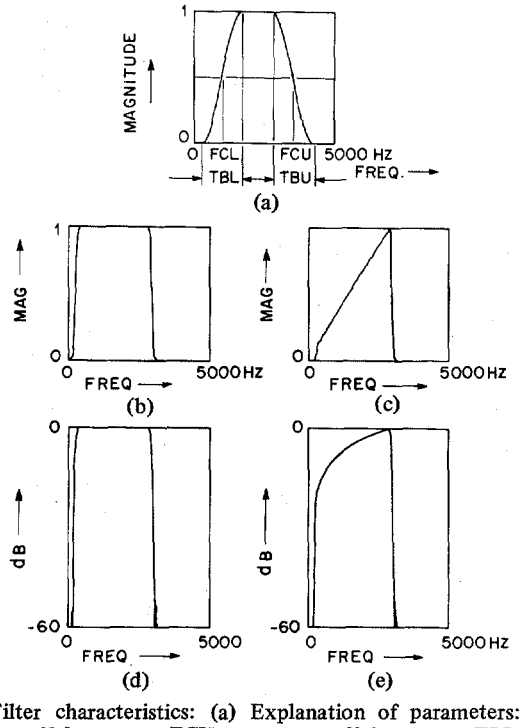

Fig. 2. Filter characteristics: (a) Explanation of parameters: FCL = lower cutoff frequency; FCU = upper cutoff frequency; TBL = lower transition band; TBU = upper transition band. Transition characteristic is a half-sine in each band. (b) Band-limiting filter. FCL = 300, FCU = 3000, TBL = 200, TBU = 200. Linear magnitude scale. (c) Band-limiting filter with superimposed 6 dB/octave preemphasis. Linear magnitude. (d) Band-limiting filter as in (b) but logarithmic magnitude. (e) Band-limiting preemphasizing filter as in (c) but logarithmic magnitude.

For each process the speech was band limited to 300-3000 Hz and preemphasized. Twenty references, 10 talkers, and 30 utterances per talker were used in each comparison (actually 29 utterances for one of the clean speech talkers). The results are given in Fig. 4 show that there is negligible practical difference between OLPA and OLPB. The small differences consistently favor OLPB, in respect of mean distance D_M and optimum mix distance D.

OLPA does not incorporate the off-diagonal distance D_{OFF}, and that component was not included in the results shown in Fig. 4. Optimum distances produced with OLPB including D_{OFF} were also investigated, but the results are not significantly different. Similarly, omission of D_V, retaining D_{OFF},

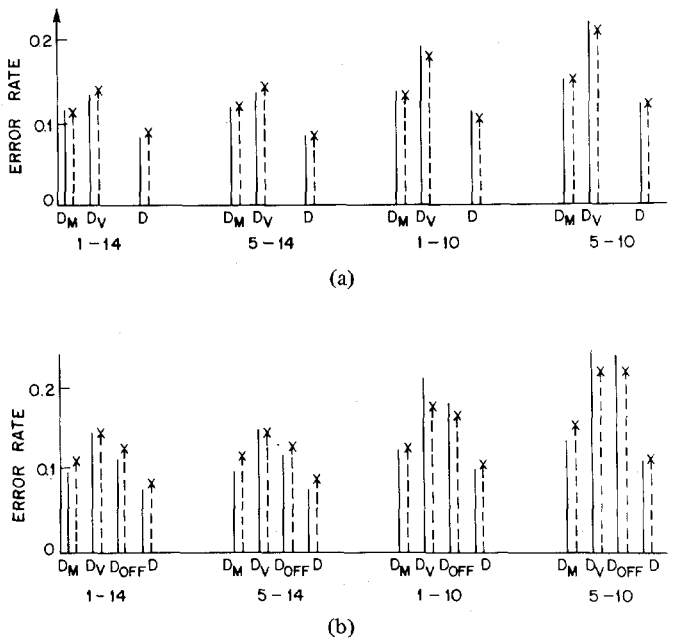

Fig. 3. Effect of 6 dB/octave preemphasis: with preemphasis (——); without preemphasis (----); (a) with OLPA; (b) with OLPB. The numbers, e.g., 1-14, indicate summing range of orthogonal parameters. In each summing range group, the dashed bars refer to D_M, D_V, D_{OFF}, and combined D, from the left. The full bars paired with the dotted bars have the same significance, there being no D_{OFF} associated with OLPA.

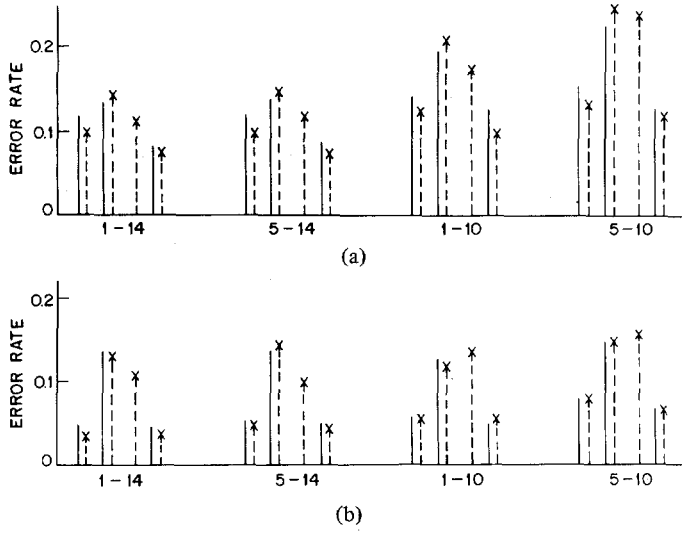

Fig. 4. Comparison of OLPA and OLPB: OLPA (----), OLPB (——). The numbers, e.g., 1-14 indicate summing range of orthogonal parameters. In each summing range group, the dashed bars refer to D_M, D_V, D_{OFF}, and combined D, from the left. The full bars paired with the dotted bars have the same significance, there being no D_{OFF} associated with OLPA. (a) results for telephone speech; (b) results for clean speech.

yields verification rates which are not significantly different. These observations refer to the use of these distances in conjunction with the reflection (parcor) coefficients; D_V and D_{OFF} have major roles in FLS described later.

A. *Summing Range*

The earlier work [1] had suggested that all the information relevant to verification is retained, if the limits of summation N_1 and N_2 in (13) are 5 and 10. Omitting contributions of parameters 1-4, i.e., including 5-14 does give negligible change as compared with the full range 1-14. (Fig. 4) There is a clear deterioration when those in the range 11-15 are omitted and a larger deterioration when both 1-4 and 11-15 are omitted, leaving 5-10. From these observations, it appears that useful verification information is included in the ranges 1-4 and 11-15, but it remains to be seen whether other advantages may be obtained by omitting either or both of these ranges. In Section VII-B we study the proposition that exclusion of some of the parameters reduces sensitivity to variations of transmission.

VI. Augmentation of Observation Vectors

The distances D_M, D_V, and D_{OFF} make use of particular aspects of the available data, namely, the variances and covariances of the elements of the data vectors and their means. The covariances are equivalent to estimates of linear dependence of the elements and would not be sufficient to achieve the potentials of the curved dividing surface shown in Fig. 5.

Further, it was thought that dependence between data observed at different times separated by particular delays might be a characteristic relevant for identification; this could correspond, for example, to a characteristic dynamic behavior of some articulator(s).

Consequently, some experiments were made using data vectors augmented by various, fairly arbitrary, combinations of delayed and nonlinearly transformed versions of the original data. The augmented vectors had 31 elements.

Results are given in Table III. It is seen that all the results for D_M and the composite distance D are at least as good as those for the original OLPB. In most cases, the improvements are negligible, however. The main point of interest is that the nonlinear combination yielded a D_M result (0.074) significantly better than that for OLPB (0.095), and effectively as good as that for the composite D (0.073) obtained with OLPB. The advantages available via the nonlinear processes and via the variance and covariance distances do not appear to be cumulative.

As there did not appear to be promise of significant performance or computational advantage in pursuing the extremely wide range of arbitrary possibilities in this area no further exploration was made.

VII. Variations of Transmission

In conventional telephone transmission systems major perturbations of the speech waveform are caused by

1) amplitude versus frequency characteristics,
2) phase versus frequency characteristics,
3) carrier-frequency offset, contributing time-varying phase, and
4) additive random noise, including various types of quantizing noise.

Additional less significant perturbations are caused by nonlinearities (largely in the transducers), echoes, voice switched devices, impulsive and systematic noise (such as tones), room noise, and room reverberation.

Transmission systems for special purposes such as bandwidth compression and storage compression make perturbations

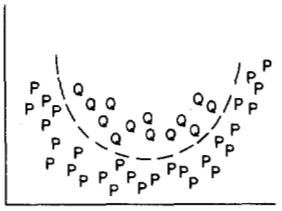

Fig. 5. Possibility of nonlinear transformation aiding separability of populations of P and Q.

TABLE III
ERROR RATES WITH AUGMENTATION OF OBSERVATION VECTOR

Process	D_M	D_V	D_{OFF}	D
OLPB ($N_1 = 1, N_2 = 14$)	0.095	0.143	0.111	0.073
Time-shifted obs.	0.090	0.175	0.123	0.067
Products of delayed obs.	0.079	0.177	0.106	0.072
Nonlinear combinations	0.074	0.166	0.150	0.073

which are harder to characterize. However, virtually all of these preserve the general features of the short-term spectrum as evidenced in spectrograms.

The statistical talker verification systems discussed in this paper all use only information about the short term autocorrelation or, equivalently, the short term power spectral density. The less significant perturbations in conventional systems do not have serious effects on the spectrum, except perhaps, for particular kinds of noise which occur rarely and are likely to be recognized and taken into account by a user. Examples would be vehicle noise, steady tones, and bursts of switching noise. The major perturbations are likely to affect the short-term spectrum primarily via 1) the amplitude versus frequency characteristics. Phase or time delay effects, 2) and 3), are orders-of-magnitude too small to cause serious changes in the short-term spectrum, and random noise is usually of such small power that it does not mask significant features of the spectrum. However, it is probably worth considering as an experimental variable, for safety.

If all telephone connections imposed the same amplitude versus frequency characteristics they would not be a potential source of confusion in the comparison of speech of various talkers. However, major variability does arise from the variety of telephone instruments, lengths of local line, and presence or absence of carrier systems.

Because carrier systems in the Bell System have carefully controlled characteristics they may be regarded as band limiting filters, and thus, suitable band limitation in a talker verification system should remove variability due to presence or absence of carrier systems. Consequently, all speech used in the experiments was band limited with the characteristics shown in Fig. 2(c). As noted in Section IV all speech was preemphasized except where it is stated otherwise.

Examples of the characteristics introduced by various telephones are shown in Fig. 6 [10]. The variability and departures from the general trend in these characteristics could be described by a moderately small number of poles and zeros, having amplitudes and depths of a few dB about the trend. Consequently, the effects of simple peaks and simple dips (Fig. 7) were investigated. These perturbations were created

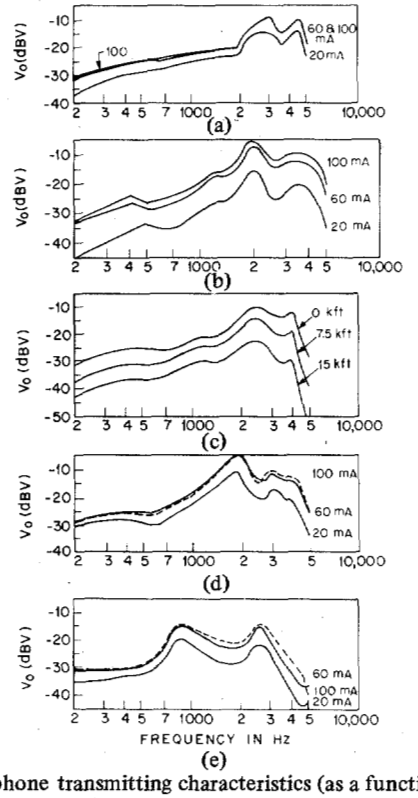

Fig. 6. Telephone transmitting characteristics (as a function of line current except as noted): (a) 500 telephone set. (b) 300 telephone set. (c) Slimline telephone set (as a function of line length). (d) Ericophon. (e) Japanese DO-8.

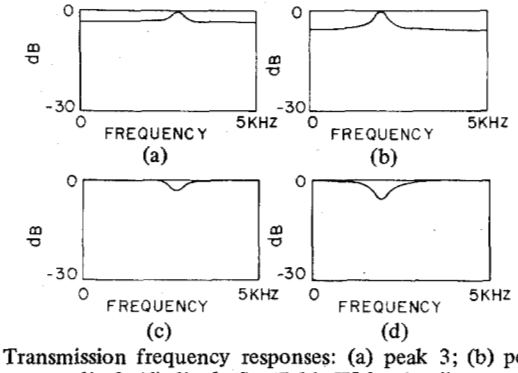

Fig. 7. Transmission frequency responses: (a) peak 3; (b) peak 6; (c) dip 3; (d) dip 6. See Table IV for details.

TABLE IV
TRANSMISSION PEAKS AND DIPS

	Frequency Hz	Pole b.w. Hz	Zero b.w. Hz	Resultant amplit., dB
Peak 3	2700	200	300	3.5
Dip 3	2700	300	200	3.5
Peak 6	2000	200	400	6.0
Dip 6	2000	400	200	6.0

by pole-zero pairs as in Table IV. The labels of the conditions, e.g., peak 3, are related to the amplitude of the perturbations.

Clean speech was used to avoid artifacts, and all utterances were processed by band limiting and preemphasis. To study the effect of a difference in transmission between the condition used for making a reference and the condition for transmission of a test utterance, references were produced using

only peak-transmission, and then only dip-transmission. These were used for comparisons with the unmodified test utterances.

Before proceeding to describe the results, a new process which was devised to overcome shortcomings which became apparent in initial peak tests of the OLP processes is introduced.

A. Filtered Logarithmic Spectrum: Process FLS

During testing of OLPA and OLPB, it became apparent that the full promise of independence of the transmission media was not borne out. Considerations of how to obtain independence led to the idea of time differences of logarithmic spectral density. The differencing would ensure that only spectral variations are included rather than absolute values which would be influenced by the medium. The logarithm would put equal proportionate variations in different parts of the spectrum on an equal basis, removing coloration in the medium.

Consequently, the observation vector was made up with

$$x_i = \log p_{i,n} - \log p_{i,n-1}, \quad i = 1, \cdots, 14 \tag{22}$$

where $p_{i,n}$ is the energy in frequency band i at frame n. Initial results were poor and this fact was attributed to the fluctuations in p_i due to the relationship between computation window and pitch period. Such an effect can be appreciated by considering the pitch bars in a broadband spectrogram.

A partial accumulation of the x_i of (22) smoothed this effect and gave promising results. The process was thus

$$x_{i,n} = ax_{i,n-1} + \log p_{i,n} - \log p_{i,n-1}. \tag{23}$$

This corresponds to a low-pass filtering of $\log p_{i,n}$ by a filter of z transform $H(z)$:

$$H(z) = \frac{1 - z^{-1}}{1 - az^{-1}}. \tag{24}$$

The implementation used $a = 0.9$, which with the frame interval of 10 ms gave a first order system with time constant 0.095 s and 3 dB cutoff at 1.68 Hz. The value was chosen as a reasonable compromise between averaging time and response to changes.

The spectral analysis was performed by using 256-sample data records, windowed by a Hamming window and Fourier transformed, giving frequency components at $10000/256 \doteq 39$ Hz intervals. The resultant real and imaginary components, A_k and B_k, respectively, were squared and added in groups of 6 to give the energy values p_i:

$$p_i = \sum_{k=k_i}^{k_i+5} (A_k^2 + B_k^2) \quad i = 1, 2, \cdots, 14 \tag{25}$$

where $k_i = 6, 12, 18, \cdots, 84$. Thus, the frequency range from about 230–3280 Hz was covered by the 14 groups. Signal information outside the nominal 300–3000 Hz band was effectively eliminated by the band-limiting preprocessing filter. The analysis was repeated at 10 ms intervals.

The spectrum analyzer parameters were not optimized; they represented objectives of keeping the analysis window and repetition interval comparable with those for the OLP processes, providing 14 coefficients, and convenience. There were not persuasive reasons for doing otherwise.

The resultant method, described by (23) and (25) will be referred to as the filtered log spectrum (FLS).

B. Effects of Variation in Transmission

Table V shows the error rates obtained for clean speech with various transmission conditions imposed on the speech signals used to make the references. The term "flat" means that no distortion was used. The results are given for three verification systems, OLPA, OLPB, and FLS, and the column N_2 refers to the upper limit of the range of summation of the orthogonal parameters. The OLPA and OLPB systems, it will be recalled, are substantially the same in performance. The discrepancy between the results given for the Flat OLPA and OLPB systems is believed to be attributable to a slightly different selection of utterance files being available for the two experiments, due to a deterioration of digital data tapes.

The main contribution to the discrimination for the OLP processes is via D_M. The columns headed D contain the error rates contained by a near-optimum sum of D_M, D_V, and D_{OFF} where relevant. For the OLP processes D_V and D_{OFF} usually do not make much difference.

For all perturbations of the reference speech both OLP processes show significant deterioration of the D_M-based error rates. In all but one case (dip 3 with OLPA) the error rates based on D are also distinctly inferior when the perturbations are present.

In contrast, the results for the FLS process are very stable. To demonstrate the transmission effects better Fig. 8 shows the relevant distributions. We see that the dominant effect for the OLP processes in the region of equal error rates is modification of the distribution of same-distances. In contrast, the distributions for the FLS process are virtually unchanged by the imposition of the transmission perturbations.

The deterioration of the error rate, and the shift of the distribution of same distances for the OLP processes occurs substantially equally for both ranges of summation of the orthogonal parameters, i.e., 1–14 and 1–10. This fact is not consistent with the proposition that parameters 11–14 accommodate variations in transmission while leaving the verification ability unchanged. Furthermore, with the exception of dip 3 with OLPA, the reduction of N_2 from 14 to 10 is always associated with a deterioration of the error rates for D_M and D_{OFF}.

VIII. Effects of Reference Composition

Three aspects of the composition of the reference files (which contain E, Λ, and ϕ) were investigated.

1) The effect of omitting the last term in (27) in the averaging of the covariance matrices of the several utterances to make the reference covariance matrix, in OLPA (Appendix I).

This was studied by assembling references by both methods, and using the distance calculation of OLPB in conjunction with both, in separate runs. The differences were negligible.

2) The effect on the distances of the number of utterances and, hence, the number of frames on which a reference is based. This is pertinent to the comments in Section II-B.

Fig. 9 shows the effect of changing the number of utterances included in a reference from 5 to 10. For OLPA we see that

TABLE V
ERROR RATES FOR VARIOUS TRANSMISSION CONDITIONS

Transmission Condition (References)	N_2	Verification System: OLPA D_M	OLPA D_V	OLPA D	OLPB D_M	OLPB D_V	OLPB D_{OFF}	OLPB D	FLS D_V	FLS D_{OFF}	FLS D
Flat	14	0.046	0.131	0.045							
Peak 3	14	0.055	0.113	0.051							
Dip 3	14	0.060	0.164	0.058							
Flat	10	0.054	0.122	0.047							
Peak 3	10	0.077	0.135	0.060							
Dip 3	10	0.064	0.113	0.042							
Flat	14				0.058	0.145	0.112	0.058	0.115	0.062	0.061
Peak 6	14				0.087	0.149	0.140	0.080	0.117	0.061	0.060
Dip 6	14				0.095	0.187	0.155	0.094	0.121	0.065	0.063
Flat	10				0.065	0.123	0.145	0.063			
Peak 6	10				0.108	0.188	0.213	0.100			
Dip 6	10				0.106	0.178	0.208	0.094			

the distribution of distances D_M moves to the right an amount corresponding approximately to doubling each distance. This is consistent with the expected behavior of D_M, the OLPA distance as given in (30). In contrast, D_M for OLPB is substantially unchanged.

Inspection of the distributions (Fig. 9) for OLPB and FLS shows that the main effect in the region of equal error rates is the left shift of the distribution of self-distances, which is consistent with the 10-component references lying more centrally in the space of the relevant population. This effect is examined in more detail in [6], for D_V and D_{OFF}. We notice also, that the distributions for different-distances tend to move slightly to the left, for larger distances. Presumably, this effect is attributable also to the greater consistency of the 10-component references.

Slight discrepancies might result in this experiment because changing the number of utterances in a reference may change the reference values of E, Λ, and ϕ, as well as $\bar{J}_m$. Also, adding utterances to a reference depletes the number of utterances available to compare with the reference, because utterances included in the reference are excluded from comparison with that reference.

3) The effect on the error rate of the numbers of frames (or utterances) on which a reference is based. Practically, this aspect is the most relevant and it is associated with the extent to which a reference adequately represents its parent population.

Table VI compares the error rates for 5- and 10-utterance references for OLPA, OLPB, and FLS. For each process and each distance the error rates are significantly lower for the 10-utterance references.

IX. IMPOSTORS

In addition to the utterances of the talkers in the main set, the 40 utterances from the 40 impostors were compared with all the 10 component references, there being 2 references per talker, i.e., 20 of these. Both OLPB and FLS were used. The values of decision threshold distance D_T corresponding to equal error rates in the main 10-component-reference trials were used.

The distributions obtained for the impostors with OLPB are shown superimposed dotted on the distributions obtained in the main comparison in Fig. 10. Using the same D_T, the error rate for the acceptance of impostors is about 0.1, i.e., impostors would have been accepted on about 80 of the 800 comparisons. The expected number is 0.055 (the original main test equal error rate) $\times$ 800 = 44, with a standard deviation of just under 7. Thus, the difference between the main talker error rate and the impostor acceptance rate is highly significant statistically, but in view of the observations concerning population inhomogeneity, it is not valid to draw any firm conclusion. Most of the impostor acceptances were in comparisons with only two talkers.

In contrast, when impostors were tested with FLS the distance distributions were virtually indistinguishable from those obtained in the main test, and the impostor acceptance rate was substantially the same as the error rate in the main test. Again, because of the small numbers it is not valid to draw firm conclusions, although it is tempting to do so.

X. CONCLUSIONS

The full promise of orthogonal linear prediction talker verification has not been supported in that variations in the transmission medium have been found to cause significant variations in the resultant distances and error rates. It is true, however, that the contributions to accuracy made by parameters 11-14 are very small in the absence of transmission perturbations.

The covariances of the observed coefficients were found to give useful information for classification independently of the mean values of the coefficients. The theory of this approach to pattern recognition is given in [6]. The covariance distance measure and the need to reduce the effects of transmission distortions led to the invention of the method of filtered logarithmic spectrum, in which the short-time spectrum of the received signal, as determined by a bank of 14 filters, was used. The absolute values of the spectrum which are influenced by transmission coloration are eliminated by taking logarithms and high-pass filtering. This method proved robust against 6 dB peaks and dips imposed in the transmission path and is believed to be the major contribution of the present paper.

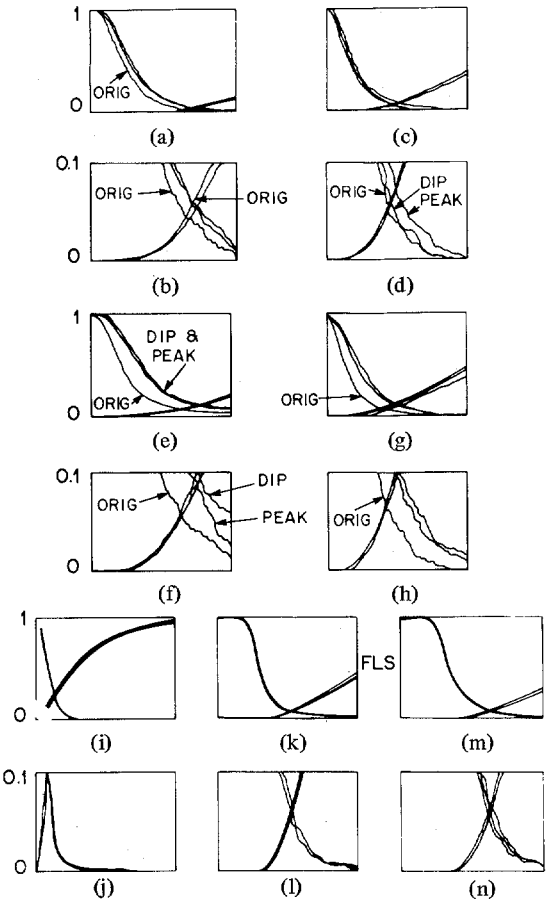

Fig. 8. Effects of transmission perturbations on distributions of distances. Vertical axes are probabilities of error.

Graphs	Process	Transmission		N_2	Distance
a	OLPA	peak 3	dip 3	14	D_M
b	"	"	"	14	D_M
c	"	"	"	10	D_M
d	"	"	"	10	D_M
e	OLPB	peak 6	dip 6	14	D_M
f	"	"	"	14	D_M
g	"	"	"	10	D_M
h	"	"	"	10	D_M
i	FLS	"	"	14	D_V
j	"	"	"	14	D_V
k	"	"	"	14	D_{OFF}
l	"	"	"	14	D_{OFF}
m	"	"	"	14	D
n	"	"	"	14	D

The number of utterances needed to define a reference which is typical of a talker is a key design feature. It was found that references based on 10 utterances gave significantly better results than references based on 5 utterances. Equal-error rates of about 5 percent appear to be readily attainable.

Subsidiary investigations were made of some other aspects.

1) Preemphasis of 6 dB/octave which is commonly used in association with LPC analysis gave a very slight advantage for the LPC systems and was then used regularly in the absence of any counter indications.

2) The details of OLPA as given in [1] differed slightly from what might be described as correct statistical procedures. However, the practical effects were negligible except that OLPA was unnecessarily sensitive to the number of frames used in defining a reference.

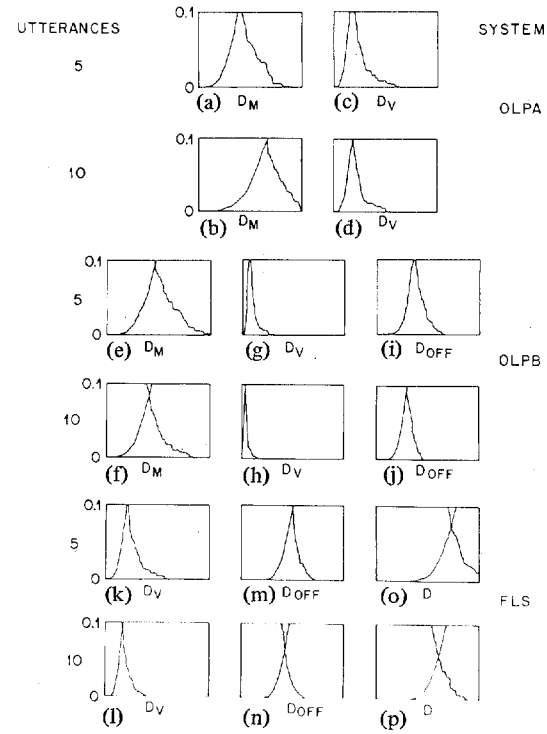

Fig. 9. Effects of numbers of utterances in a reference on distributions of distances. All vertical scales are 0 to 0.1 error rate; all horizontal distance scales are 0 to 200.

3) In the statistical design of verification experiments two factors contribute to the confidence limits, namely, the variations to be expected in a homogeneous population and lack of homogeneity. The latter is a potentially serious factor as a small number of talkers contribute a large proportion of errors. It is possible that further investigation of the effects of reference size may ameliorate this effect to some extent.

4) We looked for improvements which might be available through the augmentation of the observation vectors by values derived from the observations. Nonlinear combinations of observations, time-shifted values, and products of time-shifted values were tried. Only marginal improvements were found, and these were not at all commensurate with the increase in computation involved. It is not conclusively demonstrated that valuable augmentations do not exist, but the robustness against improvement for those forms tried suggests strongly that further exploration in this direction is unlikely to be profitable.

Further work as suggested by analysis of the results of the experiments described includes: study of the effects of reference sizes and self-variability of exceptional talkers; the effect of the lengths and nature of utterances; optimization of the parameters of the FLS process; extension to transmission perturbations other than coloration; and combination of the covariance distance measures with other independent distance measures.

Apart from the extraction of coefficients from the speech signal, which is a very simple process for FLS, the main computational load is in finding and using the eigenvalues and eigenvectors. One approach by Furui and Itakura [8] is to use eigenvalues and eigenvectors derived from averages over many talkers. Such a procedure would reduce the computation re-

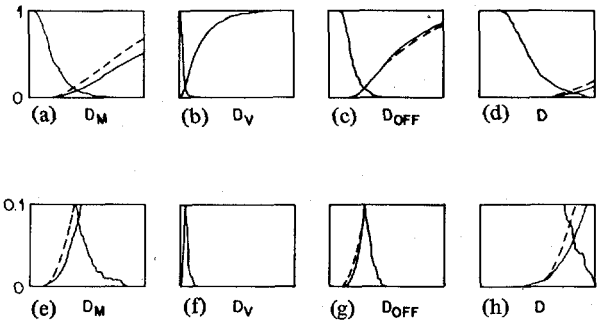

Fig. 10. Distributions of imposter distances (dotted) superimposed on distributions for main comparison OLPB, 10 component references. (a) D_M. (b) D_V. (c) D_{OFF}. (d) D. (e), (f), (g), and (h) are the same.

TABLE VI
COMPARISON OF ERROR RATES FOR REFERENCES COMPRISING 5 AND 10 UTTERANCES. TELEPHONE SPEECH

Number of utterances/reference	OLPA			OLPB				FLS		
	D_M	D_V	D	D_M	D_V	D_{OFF}	D	D_V	D_{OFF}	D
5	0.115	0.131	0.080	0.095	0.143	0.111	0.073	0.114	0.091	0.073
10	0.097	0.103	0.056	0.084	0.109	0.097	0.051	0.092	0.064	0.057

quired in updating references, but would not eliminate the transformation associated with each test utterance.

Li and Hughes [9] used covariances without such transformation and it would be worth testing their procedures, which were, more or less, correlations of the covariance matrices of the reference and test utterances.

APPENDIX I

DIFFERENCES BETWEEN METHODS OF CALCULATION

A. Calculation of C_{ij}

The C_{ij} referred to in (14) and (15) are normally calculated in the form

$$C_{ij} = \frac{1}{N} \sum_{n=1}^{N} x_{in}x_{jn} - \bar{x}_{in}\bar{x}_{jn} \tag{26}$$

which may be separated as

$$\begin{aligned} C_{ij} &= \frac{1}{N} \sum_{l=1}^{L} N_l \left\{ \frac{1}{N_l} \sum_{n=1}^{N_l} x_{iln}x_{jln} \right\} - \frac{1}{N} \sum_{l=1}^{L} \sum_{n=1}^{N_l} \bar{x}_{in}\bar{x}_{jn} \\ &= \frac{1}{N} \sum_{l=1}^{L} \sum_{n=1}^{N_l} x_{iln}x_{jln} \\ &\quad - \frac{1}{N} \sum_{l=1}^{L} \sum_{n=1}^{N_l} (\bar{x}_{iln} - \Delta_l)(\bar{x}_{jln} - \Delta_l) \\ &= \frac{1}{N} \sum_{l=1}^{L} \sum_{n=1}^{N_l} x_{iln}x_{jln} \\ &\quad - \frac{1}{N} \sum_{l=1}^{L} \sum_{n=1}^{N_l} \bar{x}_{iln}\bar{x}_{jln} \\ &\quad + \frac{1}{N} \sum_{l=1}^{L} \sum_{n=1}^{N_l} [(\bar{x}_{iln} + \bar{x}_{jln})\Delta_l - \Delta_l^2], \end{aligned} \tag{27}$$

where $\Delta_l = \bar{x}_{iln} - \bar{x}_{in}$, i.e., the difference between the mean for x_{in} in utterance l, and the mean for all elements in the L utterances.

The first two terms of C_{ij} in (26), i.e.,

$$C_{ij} = \frac{1}{N} \sum\sum x_{iln}x_{jln} - \frac{1}{N} \sum\sum \bar{x}_{iln}\bar{x}_{jln} \tag{28}$$

will be recognized as giving a value of C_{ij} which is the average of C_{ijl},

$$C_{ijl} = \frac{1}{N} \sum_{n=1}^{N_l} (x_{iln} - \bar{x}_{iln})(x_{jln} - \bar{x}_{jln}), \tag{29}$$

C_{ijl} being the covariance estimate for utterance l. The "weighted average of the calculated covariance matrices in the design set," (9) in [1] corresponds to this average, i.e., C'_{ij}, rather than the true covariance C_{ij}, (15) of the set of utterances used for the reference.

B. Calculation of D_M

Sambur [1] calculated a distance d_m, similar to D_M in (10) as

$$d_m = \sum_{i=N_1}^{N_2} \left(\frac{\bar{\phi}_{im} - z_i}{\sqrt{\lambda_{im}}} \right)^2 \bar{J}_m. \tag{30}$$

Here N_1 and N_2 denote the summing limits, adjusted to include only terms valuable for discrimination between talkers and to exclude terms believed to be more characteristic of the text or of the medium; $\bar{\phi}_m$ was the reference value of ϕ_i for the mth talker and z_i was the corresponding transformed observation value. The reason for inclusion of $\bar{J}_m$ in d_m was not given and it is not required by the probability formulation, (7) and (8).

The anomalous nature of d_m as given by (30), Appendix I

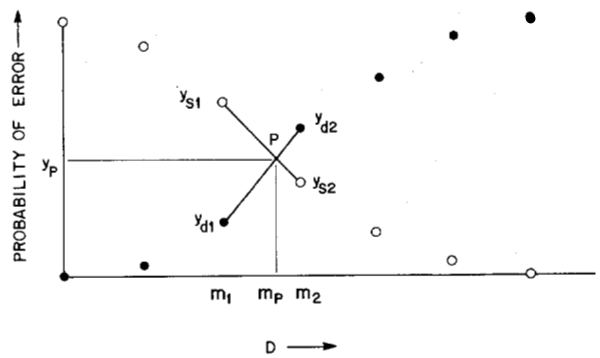

Fig. 11. Determination of crossover point P which is taken to define equal probabilities for errors of each class.

is evident if we consider two nominally similar references each based on a large number of utterances from the same talker; the numbers of frames being $\bar{J}_m$ and $2\bar{J}_m$. Because of the large number of utterances in each reference the $\bar{\phi}_m$ and λ_{im} are close estimates of these population parameters. Hence, the values of d_M obtained from the second reference will be close to twice the value obtained for the first reference.

Presumably, this effect did not prove critical because all the references used in [1] were of similar sizes. We report in Section VIII on an experiment to examine this effect. In a paper [6] presenting the theory of the covariance distance it is shown that the number of degrees of freedom relevant to such statistical calculations is more closely related to the number of phonemes than to the duration of an utterance.

Appendix II

Determination of Crossover Point

In Fig. 11 the points · and ⊙ represent the distributions of different and same-distances, respectively, classified into discrete ranges and normalized. The equal-error point P is calculated to lie at the intersection of the linear interpolations of the values y_{d1}, y_{d2}, y_{s1}, and y_{s2} in the relevant bins, numbers m_1 and m_2. By simple algebra we find

$$y_P = \frac{y_{d2}y_{s1} - y_{d1}y_{s2}}{y_{s1} + y_{d2} - y_{s2} - y_{d1}} \tag{31}$$

and

$$m_P = m_1 + \frac{y_{s1} - y_{d1}}{y_{s1} + y_{d2} - y_{s2} - y_{d1}}. \tag{32}$$

Acknowledgment

It is a pleasure to acknowledge helpful discussions with many colleagues at Bell Laboratories, especially with B. S. Atal, J. L. Flanagan, S. E. Levinson, L. R. Rabiner, and A. E. Rosenberg. Two anonymous referees made valuable contributions with regard to presentation.

References

[1] M. R. Sambur, "Speaker recognition using orthogonal linear prediction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 283–289, Aug. 1976.
[2] K. Fukunaga, *Introduction to Statistical Pattern Recognition.* New York: Academic, 1972.
[3] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *J. Educ. Psychol.*, vol. 24, pp. 417–441, 498–520, 1933.
[4] P. Jesorsky, "Principles of automatic speaker-recognition," in *Speech Communication with Computers*, L. Bolc, Ed. New York: Macmillan, 1978.
[5] L. Fasolo and G. A. Mian, "A comparison between two approaches to automatic speaker recognition," in *Proc. 1978 IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tulsa, OK, Apr. 1978, pp. 275–278.
[6] R. E. Bogner, "Pattern recognition via observation correlations," *IEEE Trans. Pattern Anal. Machine Intell.* vol. PAMI-3, Mar. 1981, to be published.
[7] A. E. Rosenberg, "Evaluation of an automatic speaker-verification system over telephone lines," *Bell Syst. Tech. J.*, vol. 55, pp. 723–744, July–Aug. 1976.
[8] S. Furui and F. Itakura, "Talker recognition by statistical features of speech sounds," *Electron. Commun. Japan*, vol. 56-A, no. 11, pp. 62–71, 1973.
[9] K. P. Li and G. W. Hughes, "Talker differences as they appear in correlation matrices of continuous speech spectra," *J. Acoust. Soc. Amer.*, vol. 55, pp. 833–837, Apr. 1974.
[10] G. A. Ledbetter, "Objective measurements of the transmission performance and characteristics of telephone sets," Bell Labs. Intern. Rep.

Robert E. Bogner (M'74) was born in Melbourne, Australia, in 1934. He received the B.E. and M.E. degrees from the University of Adelaide, Adelaide, Australia, and the Ph.D. degree from the Imperial College of Science and Technology, University of London, London, England, in 1956, 1959, and 1973, respectively.

His doctoral research was on phase processing of angle modulated signals. He was Lecturer and Senior Lecturer in electrical engineering at the University of Queensland from 1962 to 1966 and at the College of Science and Technology, University of London from 1967 to 1973. Currently, he is Professor of Electrical Engineering and Dean of the Faculty of Engineering at the University of Adelaide. His major field is communication engineering and he has made contributions in electroacoustics, speech communication, human factors, microwave modeling, modulation, and signal processing. A subsidiary interest is in unusual electrical machines. Current efforts are in speech signal processing and ultrasonic techniques applied to automated sheep shearing. He has had a number of consultancies with industry, including two periods at Bell Laboratories.