# Speaker Recognition Using Orthogonal Linear Prediction

MARVIN R. SAMBUR

*Abstract*—Recent experiments in speech synthesis have shown that, by an appropriate eigenvector analysis, a set of orthogonal parameters can be obtained that is essentially independent of all linguistic information across an analyzed utterance, but highly indicative of the identity of the speaker. The orthogonal parameters are formed by a linear transformation of the linear prediction parameters, and can achieve their recognition potential without the need of any time-normalization procedure. The speaker discrimination potential of the linear prediction orthogonal parameters was formally tested in both a speaker identification and a speaker verification experiment. The speech data for these experiments consisted of six repetitions of the same sentence spoken by 21 male speakers on six separate occasions. For both identification and verification, the recognition accuracy of the orthogonal parameters exceeded 99 percent for high-quality speech inputs. For telephone inputs, the accuracy exceeded 96 percent. In a separate text-independent speaker identification experiment, an accuracy of 94 percent was achieved for high-quality speech inputs.

## I. Introduction

In the past few years, a great deal of research has been directed towards finding speech characteristics that are effective for automatic speaker recognition [1]-[3]. Ideally, an effective speaker recognition feature should measure some aspect of the acoustic signal that reflects the unique properties of the speaker's vocal apparatus, and contains little or no information about the linguistic content of the speech. If the selected recognition feature is indicative of both the speaker and what is being said, then the full speaker discrimination potential of the feature is only realized when the recognition process is confined to a comparison of speech samples with exactly the same linguistic component. The nontrivial and nonerror free operations of segmentation and time normalization are usually required to guarantee the success of systems that depend on such features [4].

The most effective features presently employed in automatic speaker identification systems are pitch [7], gain [7], and the linear prediction characteristics of the speech waveform [3], [7]. The first two parameters are related to the properties of the speaker's glottal source and the linear prediction parameters are indicative of the talker's vocal tract. Since these three features are by themselves sufficient to produce a high-quality synthesis of the speech utterance [6], they must necessarily contain a high degree of information about the speaker's identity. Unfortunately, these parameters are also quite obviously influenced by the exact linguistic content of the speech signal and can, therefore, not be regarded as "ideal" recognition measurements.

Manuscript received October 22, 1975; revised February 11, 1976.
The author is with Bell Laboratories, Murray Hill, NJ 07974.

In order to determine the speaker identifying properties of the speech signal, it is natural to look for guidance from the results of speech synthesis experiments. A recent experimental study has shown that by an appropriate eigenvector analysis of the linear prediction parameters, a set of orthogonal parameters are obtained that can be used to achieve a high-quality synthesis of the original utterance [5]. The interesting aspect of these orthogonal parameters is that only a small subset of the parameters demonstrate any significant variation across the analyzed utterance. The remaining orthogonal parameters are essentially constant and, for purposes of synthesis, are completely specified by their measured mean values across the utterance. Since one of the aspects of the speech that is remaining constant is the speaker, these orthogonal parameters may be indicative of the talker's identity. This hypothesis was reinforced when a subsequent experiment indicated that if the same eigenvector analysis is applied to the same utterance spoken by another speaker, the resulting mean values of the corresponding orthogonal parameters are different. The implications of these experimental results are that a set of orthogonal linear prediction parameters can be obtained for a given speaker that contain almost no linguistic information and a possible high degree of information about the speaker's identity. To verify these implications, the set of orthogonal linear prediction parameters were formally examined for their ability to differentiate speakers. The results of this examination are presented in this paper. Before presenting the details of the experimentation, we shall review the concept of orthogonal linear prediction.

## II. Orthogonal Linear Prediction

In the field of speech research, the term linear prediction refers to a highly successful representation of the speech signal $\{s_n\}$ as the output of an all-pole filter $H(z)$ that is excited by a sequence of pulses separated by the pitch period for voiced sounds, or pseudorandom noise for unvoiced sounds [6]. This model implies that within a frame of speech the output speech sequence is given by

$$s_n = \sum_{k=1}^{p} a_k s_{n-k} + G u_n \tag{1}$$

where $p$ is the total number of real and complex poles attributed to the vocal tract and glottal wave, $u_n$ is the appropriate input excitation, $G$ is the gain of the filter, and the $a_k$'s are the linear prediction coefficients (LPC's) that characterize the filter. Fig. 1 illustrates the frequency domain, as well as the equivalent time-domain representation of linear predic-

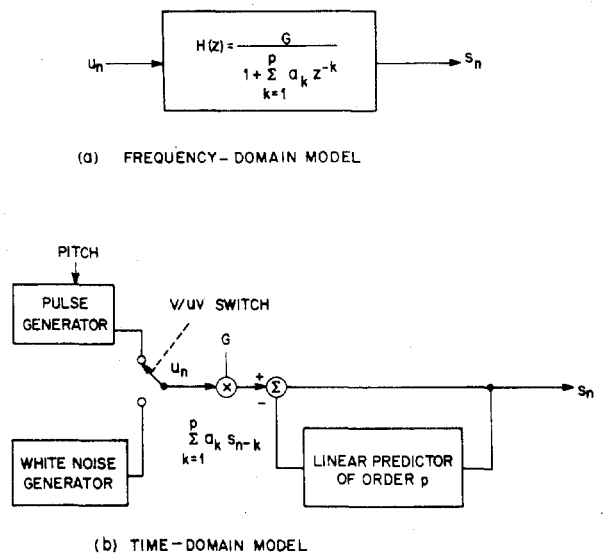


Fig. 1. Discrete model of speech production as employed in linear prediction.

tion speech production. To account for the nonstationary character of the speech waveform, the parameters $a_k$ of the modeled filter are periodically updated during successive speech frames.[1]

It is sometimes convenient to characterize the filter $H(z)$ in terms of a nonuniform acoustic tube formed by cascading $p$ uniform cylindrical sections of equal length [6]. The $p$ area coefficients, $A_i$, are uniquely related to the linear prediction coefficients by

$$\frac{A_i}{A_{i+1}} = \frac{1+\alpha_i}{1-\alpha_i}, \quad A_{p+1} = 1 \quad 1 \leqslant i \leqslant p \tag{2}$$

where the $\alpha_i$'s are termed the parcor coefficients [6]. If we denote $a_i^{(k)}$ as the $i$th LPC for a $k$th pole linear prediction model, then

$$\alpha_i = a_i^{(i)}, \quad i = 1, \cdots, p. \tag{3}$$

The area coefficients and parcor coefficients provide an alternate representation of the linear prediction characteristics of the speech wave.

The technique of orthogonal linear prediction was introduced to exploit the experimental observation that the linear prediction parameters were considerably redundant. This redundancy implies that a conventional eigenvector analysis can be used to reduce the dimensionality of the linear prediction space [8]. The eigenvector analysis involves the generation of a set of statistically uncorrelated parameters that are formed by a linear combination of the given linear prediction parameters.[2] The redundancy in the linear prediction parameters is reflected in the fact that only a small subset of the orthogonal parameters will demonstrate any significant variation across an analyzed utterance. The remaining orthogonal parameters can be effectively considered constant and only a knowledge of their particular mean values are needed to achieve high-quality speech synthesis [5].

To determine the the orthogonal parameters, we first calculate the covariance matrix, $R$ of the linear prediction parameters across the given utterance [10]. Denoting $x_{ij}$ as the $i$th linear prediction parameter in the $j$th frame, the elements of $R$ are

$$r_{ik} = \frac{1}{J-1} \sum_{j=1}^{J} (x_{ij} - \bar{x}_i)(x_{kj} - \bar{x}_k), \tag{4}$$

where

$$i, k = 1, 2, \cdots, p$$

and

$$\bar{x}_i = \frac{1}{J} \sum_{j=1}^{J} x_{ij}, \tag{5}$$

and $J$ is the number of frames in the utterance. Given the covariance matrix, the statistical variance (eigenvalue) of each orthogonal parameter is first found by solving the set of simultaneous equations [8]

$$|R - \lambda I| = 0 \tag{6}$$

where $I$ is the identity matrix, $\lambda$ is the eigenvalue, and $|Q|$ denotes the determinant of any matrix $Q$.[3] Mutually orthogonal eigenvectors are then derived as solutions of the equation[4]

$$\lambda_i \boldsymbol{b}_i = R\boldsymbol{b}_i \quad i = 1, 2, \cdots, p \tag{7}$$

where $\boldsymbol{b}_i$ is a vector with $p$ elements. In terms of the linear prediction parameters, the $i$th orthogonal parameter, $\varphi_i$, in the $j$th frame is given by

$$\varphi_{ij} = \sum_{k=1}^{p} b_{ik} x_{kj} = \boldsymbol{x}_j' \boldsymbol{b}_i. \tag{8}$$

To illustrate the behavior of the linear prediction parameters and their corresponding orthogonal parameters, Table I contains a listing of the determined eigenvalues (variances) of the calculated orthogonal parameters across the utterance "I was stunned by the beauty of the view" spoken by a typical talker. The table shows that the higher numbered orthogonal parameters have a relatively small variance and can, therefore, be considered essentially constant across the utterance for the analyzed speaker. We shall now discuss the result of our speaker recognition experiment that establish these orthogonal parameters as indicative of the speaker's identity.

## III. Speaker Recognition Experimentation (Text Dependent)

### A. Speech Data

The speaker recognition experiment involved 21 male speakers with no noticeable dialect. The speech data for the text-dependent study consisted of six repetitions of the same

[1] A frame is a segment of speech thought adequate to assume stationarity of the speech process. Typical frame lengths employed range from 10 to 30 ms.

[2] Statistically uncorrelated parameters are referred to as orthogonal parameters.

[3] Equation (6) is a $p$th-order polynomial in $\lambda$. There are $p$ solutions of this equation; $\lambda_1, \lambda_2, \cdots, \lambda_p$.

[4] A convenient Fortran program for obtaining the eigenvectors from the covariance matrix is listed in the *IBM Scientific Subroutine Package Manual* on p. 164.

TABLE I
MEASURED EIGENVALUES FOR THE ORTHOGONAL PARAMETERS DERIVED FROM THE STATISTICS OF THE LOG AREA, PARCOR, AND LPC'S OBTAINED IN A TYPICAL ANALYZED UTTERANCE

| | Log Area | Parcor | LPC's |
|---|---|---|---|
| 1 | 3.27 | 0.25 | 14.1 |
| 2 | 1.23 | 0.17 | 1.73 |
| 3 | 0.42 | 0.07 | 0.34 |
| 4 | 0.35 | 0.06 | 0.24 |
| 5 | 0.24 | 0.05 | 0.14 |
| 6 | 0.18 | 0.03 | 0.08 |
| 7 | 0.12 | 0.02 | 0.06 |
| 8 | 0.10 | 0.02 | 0.03 |
| 9 | 0.09 | 0.02 | 0.02 |
| 10 | 0.07 | 0.01 | 0.01 |
| 11 | 0.07 | 0.01 | 0.004 |
| 12 | 0.02 | 0.007 | 0.003 |

sentence spoken by each speaker on six separate occasions. The sentence analyzed was "I was stunned by the beauty of the view." The recordings were made in a quiet environment on six different days spaced over a period of three weeks.

The speech signal was sampled at 10-kHz sampling rate after bandpass filtering from 100 Hz to 4 kHz. Each utterance was then subjected to a 12th-order linear prediction analysis using the autocorrelation method to obtain the various linear prediction parameters (parcor, area, and LPC's). The analysis was conducted over contiguous speech frames of 20-ms duration. Although no time normalization was applied to the utterances, an automatic procedure was used to eliminate the consideration of silent portions and pauses in the utterance. These intervals were determined by energy and zero-crossing rate measurements [9].

### B. Recognition Procedure

In the experimental study, each set of six repetitions of an utterance for a speaker was partitioned into a design and test set. The utterances in the design set were used to evaluate each individual's reference orthogonal parameters, and the utterances in the test set were used to test the effectiveness of the orthogonal parameters. Each of the utterances rendered by a given speaker was used in turn as the test set, and the remaining five samples were used as the design set.

To evaluate the *unique* orthogonal parameters for the $m$th speaker, we first determine the reference covariance matrix for the $m$th talker. This matrix is defined as the weighted average of the calculated covariance matrices in the design set. Denoting $R_{lm}$ as the calculated covariance matrix across the $l$th utterance in the $m$th speaker's design set, we define

$$R_{\text{ref}}^{(m)} = \frac{1}{\sum_{l=1}^{L} J_{lm}} \sum_{l=1}^{L} J_{lm} E_{lm} \tag{9}$$

where $J_{lm}$ is the number of frames[5] in the $l$th utterance and $L$ is the number of utterances in the design set ($L = 5$ in this experiment). The orthogonal parameters for the $m$th speaker are then derived using $R_{\text{ref}}^{(m)}$ and (4)-(7). The $m$th speaker's orthogonal parameters are then obtained from the linear prediction parameters using the derived conversion matrix $[b_i]_m$ and (8).

In the introductory section of this paper, we noted that the least significant orthogonal parameters of a given talker are, for purposes of speech synthesis, completely specified by their mean values. Thus we can characterize the orthogonal parameters of the $m$th speaker by the conversion matrix $[b_i]_m$ and the mean values of the orthogonal parameters as determined in the design set. Denoting $\phi_{ijm}$ as the value of the $i$th orthogonal parameter for the $m$th speaker in the $j$th frame, then the average value of the $i$th orthogonal parameter for the $m$th speaker is given by

$$\bar{\phi}_{im} = \frac{1}{\sum_{l=1}^{L} J_{lm}} \sum_{l=1}^{L} \sum_{j=1}^{J_{lm}} \phi_{ijm}. \tag{10}$$

Now let us assume that we are presented with an utterance spoken by some unknown speaker and we wish to measure the dissimilarity between the unknown speaker and the $m$th speaker. Since the orthogonal parameters are by definition statistically uncorrelated, one logical measure of dissimilarity is

$$d_m = \sum_{i=p'}^{p} \left( \frac{\bar{\phi}_{im} - Z_i}{\sqrt{\lambda_{im}}} \right)^2 \bar{J}_m \tag{11}$$

where $Z_i$ is the mean value of the $i$th orthogonal parameter calculated across the utterance of the unknown talker via $[b_i]_m$ and (8), $\lambda_{im}$ is the reference eigenvalue for the $i$th orthogonal parameter of the $m$th speaker, $(p' - 1)$ is the number of leading orthogonal parameters not included in the summation, and

$$\bar{J}_m = \frac{1}{L} \sum_{l=1}^{L} J_{lm} \tag{12}$$

is the average number of frames in the utterances of the $m$th speaker's design set. This metric measures the weighted distance between the mean values of the orthogonal parameters for the presented utterance and the mean values calculated for the orthogonal parameters in the design set. The weighting factor in the distance computation is the eigenvalue of each orthogonal parameter. It can be shown that $\lambda_{im}/\bar{J}_m$ is the uncertainty or variance in the estimation of the true mean values of the orthogonal parameter from the limited design set [8, p. 345], and thus $d_m$ is just the standard distance measure for comparing uncorrelated recognition features (the features are the calculated mean values of the $m$th speaker's orthogonal parameters).

The orthogonal parameters of a given speaker can not only be specified in some sense by their mean values, but also by their variance (eigenvalue) across a given utterance.[6] It can

[5] To obtain a statistically stable estimate of the orthogonal parameters, it is advisable that the duration of the utterances be approximately 1.5 or more seconds ($J_l \geq 75$). However, for shorter durational utterances, this condition can be avoided by increasing the number of utterances in the design set ($L > 5$).

[6] Note that if we calculate the orthogonal parameters determined for speaker $A$ across an utterance of speaker $B$, there is no guarantee that the least significant orthogonal parameters will still be relatively constant for speaker $B$.

be shown that under certain broad conditions, the estimates $\bar{\phi}_{im}$ and $\lambda_{im}$ are statistically uncorrelated [8, p. 349] and thus the metric $d_m$ can be expanded to include the variance information by

$$D_m = d_m + \frac{1}{2}\sum_{i=p'}^{p}\left(\frac{V_i - \lambda_{im}}{\lambda_{im}}\right)^2 \cdot \bar{J}_m \qquad (13)$$

where $V_i$ is the measured variance of the $i$th orthogonal parameter of the $m$th speaker calculated across the presented utterance and $(\sqrt{2}/\sqrt{\bar{J}_m})\,\lambda_{im}$ is the standard deviation in the estimation of each eigenvalue [8, p. 349]. $D_m$ is again the standard distance measure for comparing uncorrelated recognition features (the features are the calculated mean values and variances of the $m$th speaker's parameters).

### C. Experimental Results

*1) Identification:* For a speaker identification evaluation,[7] the distance $d_m$ (or $D_m$) is computed for a presented utterance spoken by some unknown speaker. The unknown speaker is then identified as the $m$th speaker if $d_m$ (or $D_m$) is the smallest distance among the entire set of $d_m$'s ($D_m$'s). An error results when the actual speaker is not the $m$th talker. Fig. 2 illustrates the basic structure of the speaker identification scheme.

The identification accuracy of the orthogonal parameters was ascertained for the ensemble of 21 speakers. Since each of the six repetitions was used in turn as the test set for each of the speakers, a total of 126 judgments were made. The identification accuracy of these judgments is illustrated in Table II. The three results depicted in this table are, for the orthogonal parameters, calculated from the LPC's the log area ratio coefficients,[8] and the parcor coefficients in accordance with (4), (5), and (8). The percent identification accuracy is tabulated against the number of orthogonal parameters used to calculate the distance $d_m$ and $D_m$ $[(13 - p')$ in (11) and (12]. Thus an accuracy of 30 percent was obtained for the orthogonal parameters of the LPC's when only the least significant orthogonal parameter was used in the distance calculation ($p' = 12$) and $d_m$ was the distance metric. The identification results presented in Table II tend to saturate after the use of approximately the six or seven least significant orthogonal parameters ($p' = 7$ or $p' = 6$, respectively). This observation is consistent with the synthesis experimentation discussed in [5]. This experimentation showed that high-quality speech synthesis can be obtained with the use of only the first five or six most significant orthogonal parameters. The remaining orthogonal parameters are relatively constant and are apparently associated with the identity of the speaker as surmised by the results depicted in Table II. However, it should be noted that the last two significant orthogonal parameters did not afford a great deal of speaker identification potential. These parameters may reflect other aspects of the speech signal that are fixed. For example, they may be partially indicative of the constant frequency response of the recording apparatus and transmission system.

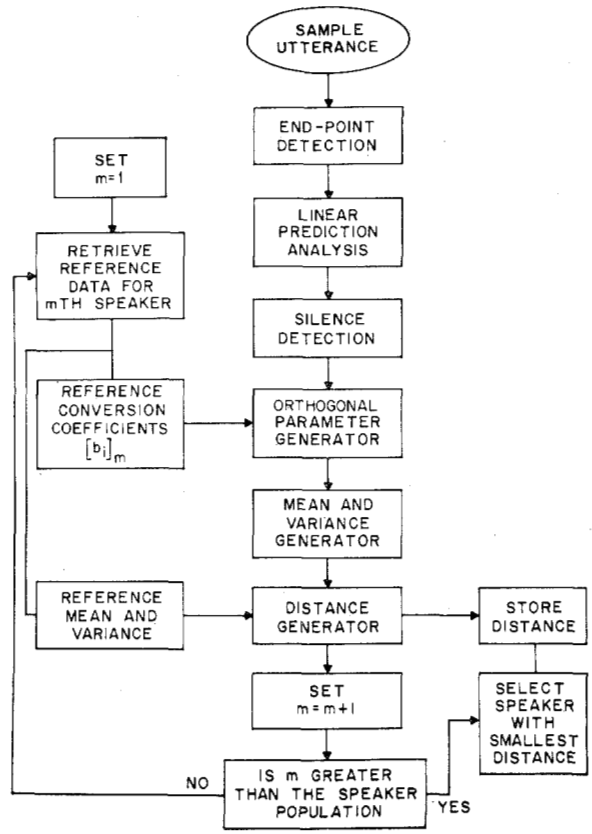


Fig. 2. Block diagram for speaker identification.

Overall, the distance metric $D_m$ achieved a higher accuracy than the metric $d_m$. In addition, the orthogonal parameters obtained from the parcor coefficients and the log area ratio coefficients outperformed the orthogonal parameters obtained from the LPC parameters. The superiority of the log area ratio and parcor orthogonal parameters can probably be attributed to their direct relationship to the physical properties of the speaker's vocal tract (see Introduction).

*2) Speaker Verification:* For speaker verification,[9] the distance between the unknown sample and the reference parameters of the claimed speaker, $m$, is compared with a threshold value. If $d_m$ (or $D_m$) is smaller than the threshold value, the speaker is verified; otherwise, he is rejected. Fig. 3 illustrates the basic structure of the verification scheme. In the speaker verification tests, one of the 21 speakers was designated as the speaker to be verified and the remaining talkers were considered impostors. Table III contains the average equal-error for the three types of parameters investigated. The equal-error rate is established by using a decision threshold that produces the same probability of incorrectly rejecting the true speaker as the probability of accepting an impostor. The performance of the orthogonal parameters for speaker verification purposes is quite similar to the results established for speaker recognition. The implication of the

[7] For a speaker identification test, the task is one of identifying the unknown speaker from among a host of possibilities.

[8] The log area ratio coefficients = $\log(A_i/A_{i+1})$.

[9] The task for a speaker verification test is to verify the identity claim of the unknown speaker.

TABLE II

| | | Number of Orthogonal Parameters | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| LPC Orthogonal Parameters | $D_m$ | 44.2 | 75.1 | 88.8 | 95.8 | 96.9 | 96.8 | 96.8 | 96.0 | 96.8 | 96.8 | 96.8 | 96.8 |
| | $d_m$ | 30.1 | 47.2 | 50.0 | 66.7 | 83.3 | 94.4 | 95.2 | 95.2 | 96.0 | 96.0 | 96.0 | 96.0 |
| Parcor Orthogonal Parameters | $D_m$ | 55.3 | 79.2 | 93.7 | 96.8 | 99.2 | 99.2 | 99.2 | 99.2 | 99.2 | 99.2 | 99.2 | 99.2 |
| | $d_m$ | 38.1 | 48.2 | 53.0 | 65.7 | 84.9 | 95.2 | 96.0 | 97.6 | 97.6 | 97.6 | 97.6 | 97.6 |
| Log Area Orthogonal Parameters | $D_m$ | 53.1 | 77.0 | 93.7 | 96.0 | 98.9 | 98.9 | 99.2 | 99.2 | 99.2 | 99.2 | 99.2 | 99.2 |
| | $d_m$ | 37.1 | 49.2 | 53.0 | 65.7 | 83.3 | 95.2 | 96.0 | 97.6 | 97.6 | 97.6 | 97.6 | 97.6 |

Speaker identification accuracy for the three linear prediction orthogonal parameters as a function of the number of least significant orthogonal parameters used for computation of distance. The metrics $D_m$ and $d_m$ are given in (11) and (12).

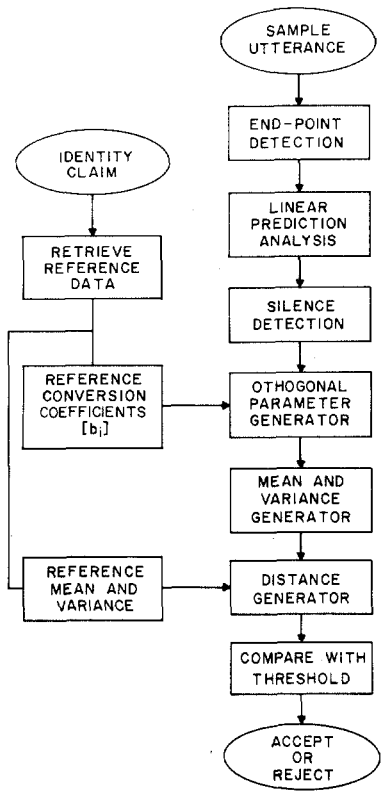


Fig. 3. Block diagram for speaker verification.

verification scores correspond to those discussed in the previous section.

## IV. Text-Independent Speaker Recognition

In many practical applications, it is desired to recognize a speaker from an utterance that does not exactly correspond to the text of the reference samples. The differences in the text of the test and the reference utterances introduce an additional variability to the problem that many automatic speaker recognition systems can not handle. In the previous section, it was demonstrated that the orthogonal linear prediction parameters could achieve a high speaker recognition accuracy without the need for any elaborate time normalization or speech segmentation schemes. Thus it may be assumed that text-independent speaker recognition can be readily achieved using the system of orthogonal parameter recognition. However, it should be appreciated that the normalization of the linguistic content of the analyzed utterance was really achieved by statistical averaging. To guarantee the success of a text-independent application of the system, the reference data must be large enough to model the overall distribution of speech sounds that are likely to occur.

To examine the potential of the orthogonal parameters for text-independent recognition, the original 21 speakers were asked to record six additional sentences. The sentences were the following:

1) We were away a year ago.
2) Every salt breeze comes from the sea.
3) I know when my lawyer is due.
4) Our yacht slid around the point into the bay.
5) May we all learn a yellow lion roar.
6) I was stunned by the beauty of the view.

The new recordings were produced on six separate occasions and processed in the manner of Section III-A.

As a first examination, the data obtained for the original analysis of the sentence "I was stunned by the beauty of the view" were used as the reference and the new sentences were used as the test data. The peak identification results using the parcor orthogonal parameters ranged from 52 to 73 percent for first five above sentences and 99 percent for the last sentence in which the text of the test and reference samples were identical.

We next expanded the reference data to include all but one of the available sentences. The remaining sentence was used as the test utterance and a total of six separate tests were made. The identification scores in this experiment ranged from 100 percent (I know when my lawyer is due) to 87 percent (Every salt breeze comes from the sea). The total results are shown in Table IV and indicate an accuracy near 94 percent. These results are an encouraging indication that text-independent speaker recognition is quite possible using orthogonal linear prediction provided the distribution of speech sounds in the reference set is not too different from the relative occurrence of sounds comprising the test set.

## V. Speaker Recognition for Telephone Inputs

With the increased use of telephone communications, it is desirable that any speaker recognition system maintain a high level of accuracy for speech inputs transmitted over telephone lines. Unfortunately, the telephone communication channel is band-limited from about 300 to 3000 Hz and much of the

TABLE III

| | | Number of Orthogonal Parameters | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| LPC Orthogonal Parameters | $d_m$ | 59.5 | 75.4 | 76.2 | 80.9 | 85.7 | 90.4 | 92.0 | 92.0 | 92.0 | 92.0 | 92.0 | 92.0 |
| | $D_m$ | 61.1 | 75.2 | 79.3 | 84.9 | 90.4 | 94.4 | 95.2 | 94.4 | 95.2 | 95.2 | 95.2 | 95.2 |
| Parcor Orthogonal Parameters | $d_m$ | 75.4 | 83.3 | 89.1 | 94.4 | 95.2 | 95.2 | 95.2 | 95.2 | 95.2 | 95.2 | 95.2 | 95.2 |
| | $D_m$ | 76.2 | 84.9 | 89.7 | 96.0 | 97.6 | 98.4 | 99.2 | 99.2 | 99.2 | 99.2 | 99.2 | 99.2 |
| Log Area Orthogonal Parameters | $d_m$ | 74.6 | 83.3 | 88.3 | 93.8 | 95.2 | 95.2 | 94.4 | 95.2 | 95.2 | 95.2 | 95.2 | 95.2 |
| | $D_m$ | 76.2 | 84.9 | 89.7 | 96.0 | 97.6 | 98.4 | 99.2 | 99.2 | 99.2 | 99.2 | 99.2 | 99.2 |

Speaker verification accuracy for the three linear prediction orthogonal parameters as a function of the number of least significant orthogonal parameters used for computation of distance. The metrics $D_m$ and $d_m$ are given in (11) and (12).

TABLE IV
TEXT-INDEPENDENT IDENTIFICATION RESULTS USING THE PARCOR ORTHOGONAL PARAMETERS AND THE METRIC $D_m$

| Number of Orthogonal Parameters | Identification Accuracy |
|---|---|
| 1 | 28.1 |
| 2 | 38.1 |
| 3 | 50.5 |
| 4 | 76.2 |
| 5 | 84.0 |
| 6 | 89.7 |
| 7 | 92.9 |
| 8 | 92.9 |
| 9 | 93.7 |
| 10 | 93.7 |
| 11 | 93.7 |
| 12 | 93.7 |

low- and high-frequency speaker identifying properties of the speech are removed. In addition, a different pair of telephone lines is typically obtained each time a new telephone call is initiated, and thus the telephone channel is slightly, but randomly, changed for each new telephone call. These problems have made speaker recognition for telephone inputs quite difficult.

To ascertain the performance of the orthogonal linear prediction approach to speaker recognition for telephone inputs, the 21 speakers each recorded six additional utterances of the sentence "We were away a year ago."[10] The utterances were transmitted over local telephone lines after band-limiting from 300 to 3000 Hz. For the transmission of each utterance, a new local telephone line was initiated. The telephone-quality utterances of the speaker were then used for a speaker identification experiment using the system of Fig. 2. For this experiment the number of utterances in the design set was again 5 ($L = 5$) and a 12th-order LPC analysis was used ($p = 12$). Using the metric $d_m$, the identification accuracy was a disappointing 87.3 percent and the accuracy using $D_m$ was an even more disappointing 83.3 percent.

After examining the data from the above experiment, it was observed that the distortions due to the telephone channel were most apparent in the distances associated with the four least significant orthogonal parameters. To understand this influence of the telephone channel, it should be noted that across a *given* utterance the characteristics of the telephone transmission media are essentially constant, and less variable than both the characteristics of the speaker and the linguistic content of the utterance. Thus, if the measured linear prediction parameters of a *given* telephone utterance are orthogonalized [from (4)-(8)], it can be expected that the first few most significant orthogonal parameters would be indicative of the linguistic content of the utterance (synthesis experiments [5]), the least significant parameter would be indicative of the telephone media, and the remaining parameters would be indicative of the speaker (recognition results for high-quality media). However, as noted above, a different pair of telephone lines may be obtained for each new telephone call and thus the characteristics of the telephone media may not be constant across the utterances in the training set. Since the orthogonalization scheme involves averaging over several utterances of a given speaker, the telephone media can no longer be represented by only the least significant orthogonal parameter. The transmission media variability spreads the representation of the telephone media across the last few significant parameters. These last few parameters are then partially indicative of the characteristics of the speaker *and* the telephone transmission media. The large distances associated with the four least significant parameters are due to the differences in the telephone media for an utterance outside the training set.

To avoid the problems of telephone media variability, it was decided to increase the order of the LPC analysis to 14 ($p = 14$). By using a 14th-order analysis, it was hoped that four poles would model the telephone media and ten poles would model the attributes of the speech. It was also hoped that an orthogonalization of the linear prediction parameters would then result in the four least significant orthogonal parameters being totally indicative of the characteristics of the telephone line. In effect, by over-specifying the number of poles needed to model the speech signal, we are leaving room for the variability of the media to be represented by the four least significant orthogonal parameters without spreading the variability across the parameters representing the speaker.

Consistent with these hopes, two modified metrics $d'_m$ and $D'_m$ were used for telephone inputs. These metrics are defined as

$$d'_m = \sum_{i=p'}^{p-4} \left( \frac{\bar{\phi}_{im} - Z_i}{\sqrt{\lambda_{im}}} \right) \bar{J}_m \tag{14}$$

and

$$D'_m = d_m + \sum_{i=p'}^{p-4} \left( \frac{V_i - \lambda_{im}}{\lambda_{im}} \right)^2 \bar{J}_m. \tag{15}$$

[10] The utterances were collected over a three-week period.

The new metrics are exactly the same as those defined in (11) and (13), but implicitly ignore the distances associated with the four least significant orthogonal parameters. For this modified system, the identification accuracy increased dramatically to 94.4 percent for the metric $d'_m$ and 96.03 percent for the metric $D'_m$. These results confirm the fact that the four least significant orthogonal parameters are indicative of the telephone transmission media. Thus the effects of the telephone channel are partially avoided by ignoring these parameters. Of course, the band-limiting effects of the phone line can never be avoided and the accuracy for this transmission media is necessarily less than the previous results obtained for high-quality media.

## VI. Conclusions

The effectiveness of the linear prediction orthogonal parameters for speaker recognition (identification and verification) was evaluated in this paper. The recognition accuracy for high-quality inputs exceeded 99 percent for the log area ratio and parcor orthogonal parameters and slightly lower for the LPC orthogonal parameters. For telephone input the accuracy exceeded 96 percent.

The important advantage of employing the linear prediction orthogonal parameters for speaker recognition is that a significant subset of these parameters are relatively constant across an analyzed utterance, and can thus be considered independent of the linguistic information and indicative of the speaker. Because of this linguistic independence, it is not necessary to time-normalize or segment the utterance to realize the full potential of the parameters. Another advantage is that the calculation of the linear prediction orthogonal parameters is an unambiguous, clearly defined operation that is easily determined from the speech signal. In addition, the linear prediction orthogonal parameters are independent of pitch and intensity information and can presumably be used to augment the recognition potential of these measurements. Finally, in view of the linear prediction orthogonal parameter's characterization of the speaker and not the linguistic content of the utterance, it may be possible to achieve text-independent speaker identification using these parameters.

## References

[1] M. R. Sambur, "Selection of acoustic features for speaker identification," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 176–182, Apr. 1975.
[2] J. J. Wolf, "Efficient acoustic parameters for speaker recognition," *J. Acoust. Soc. Amer.*, vol. 51, pp. 2044–2056, 1972.
[3] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, vol. 55, pp. 1304–1312, 1974.
[4] G. R. Doddington, "A method of speaker verification," *J. Acoust. Soc. Amer.*, vol. 49, p. 139(A), 1971.
[5] M. R. Sambur, "An efficient linear prediction vocoder," *Bell Syst. Tech. J.*, Dec. 1975.
[6] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 50, p. 637, 1971.
[7] A. E. Rosenberg and M. R. Sambur, "New techniques for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 169–176, Apr. 1975.
[8] H. Cramér, *Mathematical Methods of Statistics*. Princeton, NJ: Princeton Univ. Press, 1951.
[9] L. R. Rabiner and M. R. Sambur, "Some preliminary experiments in connected digit recognition," presented at the 90th meeting of ASA, San Francisco, CA, Nov. 1975.
[10] H. L. Van Trees, *Detection, Estimation and Modulation Theory*. New York: Wiley, 1968.

# A Comparison of Several Speech-Spectra Classification Methods

HARVEY F. SILVERMAN, MEMBER, IEEE, AND N. REX DIXON

*Abstract*—**An important consideration in speech processing involves classification of speech spectra. Several methods for performing this classification are discussed. A number of these were selected for comparative evaluation. Two measures of performance—accuracy and stability—were derived through the use of an automatic performance evaluation system. Over 3000 hand-labeled spectra were used. Of those evaluated, a linearly mean-corrected minimum distance measure, on a 40-point spectral representation with a square (or cube) norm was consistently superior to the other methods.**

Manuscript received April 30, 1975; revised January 27, 1976.
The authors are with the Speech Processing Group, Computer Sciences Department, IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598.

## I. Definition of the Problem

In our work in automatic recognition of continuous speech, a very important aspect involves classification of individual log-power spectra. These classifications are used as input to other processing stages which are intended to segment and classify at the phonemic level [1], [2]. It must be stressed that there are other alternatives for phonemic recognition, many of which do not require classification at the individual-spectrum level [3]–[10]. Preliminary, but encouraging, results with the power-spectrum classification approach have been obtained [2], [11].