TEXT-INDEPENDENT SPEAKER RECOGNITION USING ORTHOGONAL LINEAR PREDICTION

M. Shridhar, N. Mohankrishnan and M. Baraniecki*

Department of Electrical Engineering, University of Windsor
Windsor, Ontario, Canada N9B 3P4
*American Satellite Co.,20301 Century Blvd., Germantown, MD.20717, USA

ABSTRACT

The main objective of this work was to investigate the effectiveness of long-term averages of the orthogonal linear prediction parameters in text-independent speaker recognition. To investigate the possibility of feature selection, a technique using dynamic programming (1) was used to select a subset of k best features among the entire set N. The results indicate that the parameters comprising the optimal set chosen are speaker-dependent. Verification accuracies of 96.5% were obtained using the selected optimal 8-parameter (out of 12) feature set for each speaker in a verification scheme, in which the reference parameters were generated from 100 seconds of time-spaced voiced speech and the test parameters were generated from 5 seconds of voiced speech.

INTRODUCTION

Speaker recognition can be viewed in terms of two different applications (2), viz., speaker identification and speaker verification. The task in speaker identification is to associate the unknown test utterance with a particular person in the group of speakers with known speech characteristics. In speaker verification, the task is to verify if the unknown test utterance was spoken by the claimed speaker.

The basic approach to speaker recognition is to obtain a reference acoustic characterization of each speaker in the speaker population from his speech, and to store these parameters. Later on, an acoustic characterization of the test speaker's speech (either to be identified or verified) is obtained. In the case of speaker identification, the distance between the test parameters and the reference parameters of every speaker in the speaker population is calculated and the unknown speaker is identified as that reference speaker producing the smallest distance. For speaker verification, the distance between the test parameters and the reference parameters of the claimed identity are compared with a pre-determined threshold distance. If the distance is smaller than the threshold, the identity claim is accepted, otherwise it is rejected.

The purpose of this study was to investigate the suitability of using the technique of orthogonal linear prediction for text-independent speaker recognition. Sambur's (3) preliminary investigation using a rather limited text, where all the speakers read the same list of six sentences, yielded an identification accuracy of about 94%. A major contribution to the literature in the area of text-independent speaker recognition was made by Markel and Davis (4). Using a 22-feature reference parameter set composed of the mean and standard deviations of fundamental frequency and ten reflection coefficients, they obtained an identification accuracy of 98% and an equal-error verification rate of 4.25%. They used the fact that the orthogonal parameters are essentially equivalent to a linear transformation of the long-term reflection coefficient averages they had used in their study, to draw a parallel between their experimental results and Sambur's text-independent study. Based on this they came to the conclusion that a true text-independent experiment using orthogonal parameters as features, would yield results no better than about 62%.

In this work a more exhaustive study of the long-term convergence of the orthogonal linear prediction parameters, when averaged over long sections of unconstrained speech, has been carried out. Sambur (3) surmised that the least significant orthogonal parameters that exhibit small variances across the analyzed utterance, could be indicative of the talker's identity. To test this hypothesis in a text-independent speaker recognition experiment, a feature selection technique using dynamic programming (1) is used in this work to select the best k-feature subset from among the entire set N. The results of these experiments are presented in subsequent sections.

PREPROCESSING

The data base used in this study consisted of about 60 seconds of speech for each speaker, from each of six sessions, spaced at least a week apart. The six sessions spanned a period of about four months. The speaker population was made up of eight speakers. The recordings were high quality speech recordings made in a normal quiet room environment. The speech was band-limited to 4.5 kHz before sampling at a rate of 10 kHz. All the processing was done on a Data General NOVA 840 minicomputer with a high speed signal processing attachment (AP 120B). The silent and unvoiced portions of the speech were detected and removed using an energy threshold criterion. Successive 20 msec. frames were multiplied by a Hamming window and 12th order linear prediction analysis using the autocorrelation method was performed at

CH1610-5/81/0000-0197 $00.75 © 1981 IEEE

a rate of 50 frames/second.

SPEAKER RECOGNITION USING ORTHOGONAL LINEAR PREDICTION

The steps involved in determining the orthogonal parameters for speaker 'm' are as follows(3)

(a) Calculate the covariance matrix $\mathbf{R_m}$ of the LPC parameters across the given reference speech segment of speaker 'm'. If x_{ijm} is the ith linear prediction parameter in the jth frame of the mth speaker, the elements of the covariance matrix $\mathbf{R_m}$ for the mth speaker are given by,

$$(r_{ik})_m = \frac{1}{J-1} \sum_{j=1}^{J} (x_{ijm} - \overline{x_{im}})(x_{kjm} - \overline{x_{km}}) \quad (1)$$

where,

i, k = 1, 2, ..., P

J is the number of frames in the reference speech segment, P is the order of the LPC model and $\overline{x}_{im}$ is given by,

$$\overline{x_{im}} = \frac{1}{J} \sum_{j=1}^{J} x_{ijm} \quad (2)$$

(b) The set of eigenvalues λ_{im} of $\mathbf{R_m}$ are obtained by solving the set of simultaneous equations given below, for each speaker 'm'.

$$| \mathbf{R_m} - \lambda \mathbf{I} | = 0 \quad (3)$$

where $\mathbf{I}$ is an identity matrix.

(c) Mutually orthogonal eigenvectors $(\mathbf{b_i})_m$ are then obtained as solutions of the equation,

$$\lambda_{im} (\mathbf{b_i})_m = \mathbf{R_m} (\mathbf{b_i})_m \quad i = 1, 2, ..., P \quad (4)$$

(d) The orthogonal parameters are calculated as a linear combination of the LPC parameters.

$$\phi_{ijm} = \sum_{k=1}^{P} b_{ikm} x_{kjm} \quad (5)$$

where ϕ_{ijm} is the ith orthogonal parameter in the jth frame of the mth speaker.

(e) The reference orthogonal parameters for a particular speaker are obtained as the average of his orthogonal parameters calculated over all the frames in the reference speech segment.

$$\overline{\phi_{im}} = \frac{1}{J} \sum_{j=1}^{J} \phi_{ijm} \quad (6)$$

(f) The measure of dissimilarity 'd_m' between the ith speaker in the population and the unknown test speaker is given by,

$$d_m = \sum_{\text{chosen subset of orthogonal parameters}} \left| \frac{\overline{\phi_{im}} - z_i}{\sqrt{\lambda_{im}}} \right|^2 \quad (7)$$

where Z_i is the mean value of the ith orthogonal parameter calculated across the test speech of the unknown talker using $(\mathbf{b_i})_m$ and equation (5), and λ_{im} is the reference eigenvalue for the ith orthogonal parameter of the mth speaker.

This distance measure gives the weighted distance between the mean values of the orthogonal parameters calculated from the reference speech segment, and the mean values of the orthogonal parameters calculated across the test speech.

CONVERGENCE PROPERTIES OF THE COVARIANCE MATRIX

It can be seen from the previous section that a crucial step in orthogonal linear prediction analysis is to obtain a stable estimate of the covariance matrix of the LPC parameters. In order to find out what length of speech had to be included in calculating the covariance matrix of the LPC parameters, to ensure its convergence, the following experiment was performed. An initial estimate of the covariance matrix was obtained from just the first set of 200 frames from Session I of a particular speaker. One frame of speech is of length 20 msecs. Subsequently 200 more frames were added on from Session I and a new estimate of the covariance matrix was obtained from the first 400 frames of Session I. The difference between the two matrices was quantified using the Average Absolute Difference Measure (5) between two n x n matrices (t_{ij}) and (r_{ij}), defined as,

$$AAD = \frac{2}{n(n+1)} \sum_{j=1}^{n} \sum_{i=j}^{n} | t_{ij} - r_{ij} | \quad (8)$$

The AAD measure is a function of both distance and angle between the two matrices. For correlation coefficient matrices it ranges in value between 0 and 2. (Hence all covariance matrices used in this study were converted into correlation coefficient matrices.)

The procedure described above was repeated by adding on 200 frames at a time and evaluating the AAD between successive matrices differing only by 200 frames in their estimation. When all the 1000 frames of a particular session were used up, 200

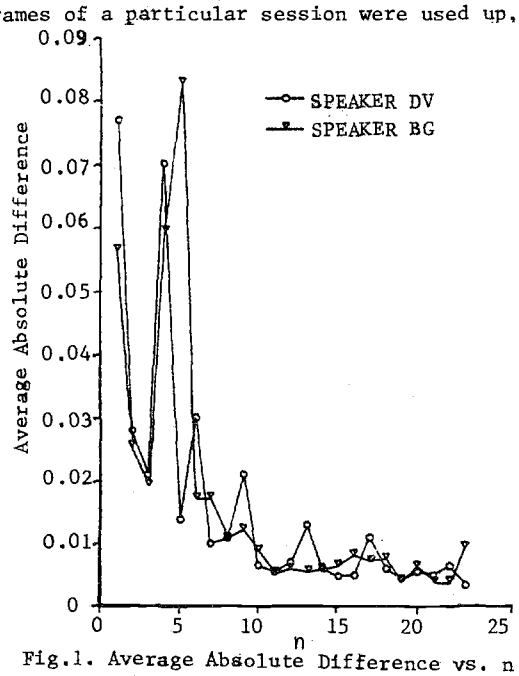

Fig.1. Average Absolute Difference vs. n (AAD at n is the AAD between two covariance matrices estimated using 200 (n+1) and 200n frames).

frames from the next session were added on in estimating the covariance matrix. The process was continued until all the 5000 frames from five sessions had been utilized in estimating the covariance matrix of the LPC parameters.

The results of the above analysis are presented in Figure 1 which shows the Average Absolute Difference between successive matrices computed as described above. Though the AAD registers jumps when frames from a new recording session are included in estimating the covariance matrix, in general it is steadily decreasing. At the end of about five sessions, it is oscillatory in nature and has decreased to a sufficiently low value for a reasonable conclusion to be drawn that about 5000 frames (100 seconds) of speech are required to obtain a fairly stable estimate of the covariance matrix.

SPEAKER RECOGNITION EXPERIMENTS

Though the results of the previous section indicate that about 5000 frames (100 seconds) of speech are needed to obtain a stable estimate of the covariance matrix of the LPC parameters, three speaker recognition experiments, in which the reference parameters were obtained from only 1000 frames, were carried out in order to show feasibility of recognition schemes even with short-duration reference texts. A fourth experiment making use of the results of the previous section was also performed. The four different speaker recognition experiments, in which the long-term properties of each speaker were characterized by a model consisting of the long-term orthogonal parameter averages, resulted from obtaining the reference speaker parameters from different portions of the speech data base and the relative choice of the test data. In all these experiments all the 12 orthogonal parameters were included in computing the distance measure.

Experiment I: The reference parameters were obtained from 1000 frames from Session I and the test data was obtained from 100 frames from Session I. Though the reference and test data were obtained from the same session, there was no overlap between the two texts.

Experiment II: The reference parameters were obtained from 1000 frames from Session I and the test data was obtained from 250 frames from Session III.

Experiment III: The reference parameters were obtained from a combination of 500 frames from each of Sessions I and II (1000 frames in all) and the test data was obtained from 250 frames from Session III.

Experiment IV: The reference parameters were obtained from a combination of 1000 frames from each of five of the six sessions (5000 frames in all) with the test data being obtained from 250 frames of the sixth session.

In all these experiments the recognition tests were repeated with multiple sets of test data to obtain statistically significant results. Table I lists the results of the above experiments. It can be seen that the recognition performance (100% for identification and 98.4% for verification) is

TABLE I: RESULTS OF EXPERIMENTS I - IV

Expt. No.	Identification Accuracy	Verification Accuracy
1	100%	98.4%
2	92.6%	95.3%
3	95.3%	95.3%
4	96.5%	96.7%

good when the reference and trial data are drawn from the same recording session (Expt.1). When the reference and trial data are drawn from different sessions (Expt.2), the recognition scores are lowered (92.6% and 95.3%). By comparing the results of Experiments 2 and 3 (92.6% and 95.3% vs 95.3% and 95.3%), the effect of generating the reference parameters from time-spaced speech data, obtained from more than just one recording session can be seen. While the identification scores show significant improvement, the verification scores are about the same. Experiment 4 which was designed based on the results of the study of the convergence of the covariance matrix of the LPC parameters, yielded an identification accuracy of 96.5% and a verification accuracy of 96.7%.

Next, the probability of feature selection among the orthogonal parameters was investigated. Recognition tests conducted with different random subsets of the orthogonal parameters used in evaluating distances, yielded some interesting results. In several instances, a feature subset made up of parameters 4 through 9 yielded better scores than the case where all the 12 parameters were included. Consequently, a systematic feature selection technique using dynamic programming (1) was implemented in order to investigate whether a subset of the orthogonal parameters would yield comparable or better results. The results obtained are discussed in the next section.

FEATURE SELECTION USING DYNAMIC PROGRAMMING

Dynamic programming is a multistage optimization technique based on the Principle of Optimality which states that whatever the initial state and decisions are, the remaining decisions must constitute an optimal policy with respect to the state resulting from the first decision. The steps in the feature selection procedure discussed below are from the work of Cheung and Eisenstein (1).

Let the set of N available features be represented by $\mathbf{X} = (x_1, x_2, \ldots, x_N)$ and let $P_n^j = (p_1^j, p_2^j, \ldots, p_n^j)$ be one of the N possible subsets chosen after n stages, where p_n^j is a feature in $\mathbf{X}$. For every x_j at the nth stage, the subset P_n^j is selected such that

$$\lambda_n(P_n^j) = \max_{x_j \notin P_{n-1}^i} D_n(P_{n-1}^i, x_j), \quad i=1,2,\ldots,N \tag{9}$$

where (P_{n-1}^i, x_j) is a subset formed by augmenting P_{n-1}^i with x_j as shown below:

$$(P_{n-1}^i, x_j) = (p_1^i, p_2^i, \ldots, p_{n-1}^i, x_j) \tag{10}$$

D_n is a chosen feature effectiveness criterion and λ_n is the maximum effectiveness measure over a collection of subsets as defined in equation (9). The optimum subset P_n at the end of n stages is obtained as shown below:

$$\lambda_n(P_n) = \max \lambda_n(P_n^j),\ j = 1,2,\ \ldots,\ N \qquad (11)$$

Dynamic programming analysis results in optimal solutions only if the k-stage problem can be broken up into k subproblems. This is guaranteed if the chosen feature effectiveness criterion, D_n, satisfies the following two conditions:

(a) D_n is a monotonic, nondecreasing function of n, the stage number.

$$D_n(P_n^j) \geq D_{n-1}(P_{n-1}^j) \quad \text{for any } p_n^j \qquad (12)$$

(b) D_n can be split into two parts, one representing the history of the process till the (n-1)st stage and the other representing the behavior of the process at the ith stage.

$$D_n(P_n^j) = f\left(D_n(p_n^j),\ D_{n-1}(P_{n-1}^j)\right) \qquad (13)$$

The feature effectiveness criterion chosen for this analysis was:

$$\text{FEC} = \sum \text{Inter-speaker distances} - \sum \text{Intra-speaker distances} \qquad (14)$$

where the distances were evaluated using the feature subset under consideration, and the summation was performed over all the distances obtained from using data from each one of the six sessions in turn as trial data. This is a logical measure to use in assessing feature effectiveness since one would like to select a feature subset that maximizes inter-speaker distances and minimizes intra-speaker distances. This feature effectiveness criterion also satisfies the two conditions listed above, thus ensuring decomposition of the k-stage problem into k subproblems.

TABLE II: RESULTS OF FEATURE SELECTION TECHNIQUE TO SELECT A SUBSET OF 6-PARAMETERS

Name of Speaker	Parameters Selected (Indicated by X)											
	1	2	3	4	5	6	7	8	9	10	11	12
BG			X		X	X	X	X	X			
BH						X	X		X	X	X	X
DV		X			X		X		X	X		X
JN				X	X	X	X		X	X		
MJ		X	X	X	X	X			X			
PA		X	X		X		X	X		X		
PL			X	X	X	X			X	X		
WM		X		X	X	X			X	X		

Table II lists the results of the feature selection procedure discussed above for choosing a 6-parameter feature set. It can be seen that the chosen parameters are speaker-dependent. These results were used in performing Experiment 5 described below:

Experiment V: This experiment was based on the same reference and trial data base as Experiment IV. In the recognition tests the speaker-dependent 6-parameter and 8-parameter feature subsets obtained through feature selection were used to compute distances.

TABLE III: RESULTS OF EXPERIMENT V

No. of Parameters in feature set	Identification Accuracy	Verification Accuracy
All 12	96.5%	96.7%
8	96.9%	96.5%
6	97%	96.3%

The results of Experiment V are presented in Table III. While the identification results show an improvment from 96.5% (all 12 parameters used) to 97% (6-parameter feature set used), the verification scores decline from 96.7% (all 12 parameters) to 96.3% (6-parameter feature set).

CONCLUSIONS

The conclusions of this investigation are as follows:

(i) Text-Independent speaker recognition using long-term averages of the orthogonal parameters as features is feasible with better than 96.5% results being obtained in recognition tests.

(ii) Speaker recognition can be achieved with as little as 5 seconds of test data with the reference parameters being extracted from about 100 seconds of speech.

(iii) The conclusions drawn by Markel and Davis (4) about the unsuitability of orthogonal parameters in a true text-independent situation do not appear to be justified.

(iv) The optimal k-feature subset selected through a dynamic programming feature selection technique are speaker-dependent. They are not made up of the least significant set of orthogonal parameters as claimed by Sambur(3).

REFERENCES

1. R.S. Cheung & B.A. Eisenstein, "Feature Selection via Dynamic Programming for Text-Independent Speaker Identification", IEEE Trans.Acoust., Speech, & Signal Processing, Vol.ASSP-26, pp. 397-403, October 1978.
2. B.S. Atal, "Automatic Recognition of Speakers from Their Voices", Proc. IEEE, Vol.64, pp.460-475, April 1976.
3. M.R. Sambur, "Speaker Recognition Using Orthogonal Linear Prediction", IEEE Trans. on ASSP, Vol.ASSP-24, pp.283-289, August 1976.
4. J.D. Markel and S.B. Davis, "Text-Independent Speaker Recognition from a Large Linguistically Unconstrained Time-Spaced Data Base",IEEE Trans. on ASSP, Vol.ASSP-27, pp.74-82, Feb.1979.
5. K.P. Li and G.W. Hughes, "Talker Differences as They Appear in Correlation Matrices of Continuous Speech,J.Acoust.Soc.Am.,Vol.55,833-7,Apr.74.