TEXT INDEPENDENT SPEAKER RECOGNITION USING ORTHOGONAL LINEAR PREDICTION

M. R. Sambur
Bell Laboratories
Murray Hill, New Jersey 07974

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 were formally tested in both a speaker recognition 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 recognition and verification, the identification accuracy of the orthogonal parameters exceeded 99%. In a separate text-independent speaker recognition experiment, an accuracy of 94% was achieved.

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.[1,2,3] 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.[4] 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 analyzed utterance. For a typical 12 pole linear prediction analysis, it has been determined that it is only necessary to transmit (along with the 7 mean values) the 5 most significant orthogonal parameters to achieve a high quality synthesis of the original utterance.[4]

The question now presents itself as to the interpretation of the 7 fairly constant orthogonal parameters. Since one aspect of the given analyzed utterance that is remaining constant is the speaker, these orthogonal parameters may be indicative of the talker's identity. This hypothesis was re-enforced 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 (as they are essentially constant across the analyzed utterance and only their mean values are needed for synthesis purposes) and a possible high degree of information about the speaker's identity (as their mean values are different for different speakers). 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.

II. RECOGNITION SYSTEM

Figure 1 illustrates the overall structure of the speaker identification* system used in our experimentation. The input utterance is initially bandpass filtered from 100 - 4 kHz and sampled at 10 kHz. After endpoint detection,[5] the utterance was then subjected to a 12th order LPC analysis using the autocorrelation method[6] to obtain the linear

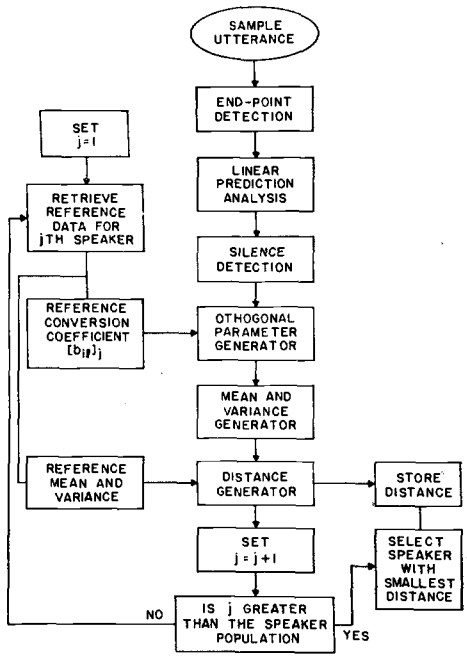

Figure 1: Block diagram for speaker identification.

* For a speaker identification system, the task is one of identifying the unknown speaker from among a host of possibilities.

prediction parameters. A silence detector[7] was used to eliminate the consideration of pauses in the spoken utterance. For each speaker the sample utterances were divided into a test set and a design set. The design set was used to calculate the propor orthogonal parameters for the given speaker and the appropriate reference statistics.

The orthogonal parameters, ϕ_i, are related to the linear prediction parameters by

$$\phi_i = \sum_{\ell=1}^{p} b_{i\ell} x_\ell = \underset{\sim}{x}^T \underset{\sim}{b}_i, \quad i = 1,2,\ldots,p$$

where p is the order of the linear prediction analysis. To determine the conversion coefficients, $b_{i\ell}$, we first calculate the covariance matrix of the linear prediction parameters, $\underset{\sim}{R}$, across the utterance. If we denote x_{ij} as the i^{th} LPC parameter in the j^{th} frame, then the elements of $\underset{\sim}{R}$ are

$$r_{ik} = \frac{1}{N-1} \sum_{j=1}^{N} (x_{ij}-m_i)(x_{kj}-m_k)$$

where

$$m_i = \frac{1}{N} \sum_{j=1}^{N} x_{ij}$$

and N is the number of frames in the utterance. The eigenvalues (or variance) of the orthogonal parameters are then found by solving the set of simultaneous equations

$$|\underset{\sim}{R} - \lambda \underset{\sim}{I}| = 0$$

where $\underset{\sim}{I}$ is the identity matrix, $|\underset{\sim}{C}|$ denotes the determinant of the matrix $\underset{\sim}{C}$, and $\lambda_1,\lambda_2,\ldots,\lambda_p$ are the eigenvalues. The conversion coefficients are then derived as solutions of the equation**

$$\lambda_i \cdot \underset{\sim}{b}_i = \underset{\sim}{R} \cdot \underset{\sim}{b}_i, \quad i = 1,2,\ldots,p$$

It is important to remember for each speaker we determine the conversion coefficients that specify his unique orthogonal parameters.

The distance metric used in our system is given by

$$D_j = \sum_{i=\ell_s}^{p} \left[\frac{(\bar{\phi}_{ij}-Z_{ij})^2}{\lambda_{ij}} + \frac{1}{2}\left(\frac{V_{ij}-\lambda_{ij}}{\lambda_{ij}}\right)^2 \right]$$

where $\bar{\phi}_{ij}$ - mean value of i^{th} orthogonal parameter of the j^{th} talker as calculated in the design set.

Z_{ij} = mean value of the i^{th} orthogonal parameter of the j^{th} talker as calculated across the unknown utterance.

λ_{ij} = standard deviation of the i^{th} orthogonal parameter of the j^{th} talker as calculated in the design set.

V_{ij} = standard deviation of the i^{th} orthogonal parameter of the j^{th} talker as calculated across the unknown utterance.

ℓ_s = starting orthogonal parameter for distance computation.

Thus if we wish to establish the distance between the unknown talker and the j^{th} speaker, we simply measure the prescribed orthogonal parameters for the j^{th} speaker across the presented utterance. We then calculate the mean and standard deviation of these parameters across the utterance and compare them with the reference data of the j^{th} speaker. The metric is simply a weighted euclidean distance using the mean and standard deviation as speaker identification features. It is important to note that these features are not dependent on time and thus the identification algorithm does not require any time normalization or segmentation.

III. EXPERIMENTAL RESULTS

Two separate experiments were undertaken to establish the speaker recognition potential of these orthogonal measurements. In the first experiment 21 male speakers made 6 separate recordings of the utterance "I was stunned by the beauty of the view". The recordings were made in a quiet environment on six different days spaced over a 3 week period. The results of a separate identification and verification† for this experiment are shown in Table 1. The results presented are for the orthogonal parameters derived from the parcor coefficients which yielded the best scores among the various linear prediction parameters examined. Thus, an accuracy of 99.2 was obtained for the parcor orthogonal coefficients when only the 7 least significant orthogonal parameters were used in the distance calculation. This result is consistent with synthesis experiments that indicated that these parameters were fairly constant across the analyzed utterance.

In a separate text independent experiment, we asked the 21 speakers to record an additional 6 sentences:

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 slide 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.

** A convenient Fortran program for obtaining the eigenvectors from the covariance matrix is listed in the IBM Scientific Subroutine Package manual on page 164.

† The task for a speaker verification system is to verify the identity claim of the unknown speaker.

Number of Orthogonal Parameters	Identification	Verification
1	55.3	76.2
2	79.2	84.9
3	93.7	89.7
4	96.8	96.0
5	99.2	97.6
6	99.2	98.4
7	99.2	99.2
8	99.2	99.2
9	99.2	99.2
10	99.2	99.2
11	99.2	99.2
12	99.2	99.2

Table I: Recognition results using the parcor orthogonal parameters for the experiments using the sentence "I was stunned by the beauty of the view."

The new recordings were produced on 6 separate occasions and processed in the same manner as experiment 1. In this experiment each of the sentences were used in turn as the test set and the remaining five sentences were used as the design set. Thus the text of the design set and the test set are different. The average identification accuracy for the 6 tests are shown in Table II. Thus even when the text of the test and the design set are different, an accuracy of about 94% can be achieved.

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

Table II: Text independent identification results using the parcor orthogonal parameters.

IV. SUMMARY

In summary, the important characteristic of the linear prediction orthogonal parameters for purposes of 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 important 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 is possible to achieve text independent speaker identification using these parameters.

REFERENCES

1. M. R. Sambur, "Selection of acoustic features for speaker identification," IEEE Trans. on Acoust., Speech, and Signal Proc., Vol. ASSP-23, No. 2, April 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. M. R. Sambur, "An efficient linear prediction vocoder," to be published in the Bell Sys. Tech. Journ., December 1975.

5. L. R. Rabiner and M. R. Sambur, "An Algorithm for Determining the Endpoints of Isolated Utterances," B.S.T.J., 54, No. 2 (February 1975).

6. J. D. Markel, A. H. Gray, Jr., and H. Wakita, "Linear Prediction of Speech Theory and Practice," SCRL Monograph No. 10, Santa Barbara, Cal., SCRL, September 1973.

7. L. R. Rabiner and M. R. Sambur, "Some preliminary experiments in the recognition of connected digits," presented at the Acoustical Society of America's 90th Meeting in San Francisco, November 1975.