A COMPARISON OF FOUR TECHNIQUES FOR AUTOMATIC SPEAKER RECOGNITION

R. E. Wohlford, E. H. Wrench, Jr., and B. P. Landell

ITT Defense Communications Division, 9999 Business Park Avenue, San Diego, CA. 92131

ABSTRACT

Four automatic speaker recognition techniques were investigated with a common speech data base to determine their effectiveness in a text independent mode. These four techniques used the correlation of short and long term spectral averages, cepstral measurements of long term spectral averages, orthogonal linear prediction of the speech waveform, and long term average LPC reflection coefficients combined with pitch and overall power.

The results of this study indicate that LPC derived parameters perform better than do those derived from cepstral and spectral data. Recognition accuracies of 95% and 93% were obtained for LPC based techniques with 13 seconds of unknown speech. The corresponding recognition accuracies for the cepstral and spectral based systems were 79% and 54% respectively.

INTRODUCTION

The purpose of this study was to test and compare four techniques of automatic speaker recognition on a common data base. Performance was compared in a text independent environment on free conversational speech. This paper details the methods used and the results obtained during the study. The four techniques compared are:

1. The correlation of short and long term spectral averages as investigated by S. Pruzansky and M.V. Mathews [1].

2. Cepstral measurements of long term spectral averages as investigated by S. Furui, F. Itakura, and S. Saito [2].

3. Orthogonal linear prediction of the speech waveform as investigated by M.R. Sambur [3].

4. Long term average LPC reflection coefficients, pitch and overall gain of the speech waveform as investigated by J.D. Markel, B.T. Oshika, and A.H. Gray [4].

These four recognition techniques were implemented on a PDP-11/60 computer and a high speed signal processor. An extremely large data base was used to determine the recognition accuracies of the four techniques. The data base consisted of eight and one half hours of speech recorded with a high signal to noise ratio and subsequently bandlimited to telephone bandwidth (300-3200 Hz).. The data base contains ten digitized interviews, each of approximately three minutes duration for seventeen talkers. All of the interviews represent free conversational speech. The interviews with each speaker were conducted once a week over a ten week period.

SPEAKER RECOGNITION TECHNIQUES

Speaker recognition as used in this paper is defined as follows. A sample of speech from an unknown speaker is analyzed and compared with data extracted from speech samples from a set of known speakers. A choice is made as to which speaker in the known set best corresponds to the unknown. In this study, the speech samples were text independent. The speakers did not all record the same material, and the recognition algorithms made no apriori use of the phonetic content of the recordings.

The four speaker recognition techniques investigated in this study can each be described as consisting of four separate functions as shown in Figure 1. The four functional blocks are described in the following paragraphs.

Speech Analysis Parameter Extraction

In the first functional block, speech analysis parameters are calculated from the raw speech. The parameters used in each of the four techniques are given in Table 1. They include filter bank parameters, cepstal coefficients, LPC coefficients, and LPC coefficients combined with pitch and power.

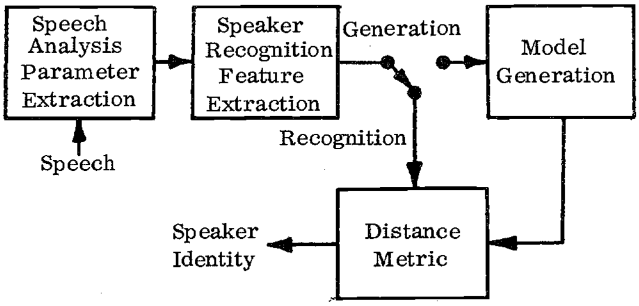
Speech Analysis Parameter Extraction
Speaker Recognition Feature Extraction
Generation
Model Generation
Speech
Recognition
Speaker Identity
Distance Metric

Figure 1. A Generalized Speaker Recognition System

CH1559-4/80/0000-0908$00.75 © 1980 IEEE

Table 1: Speech Analysis Parameters Used for Speaker Recognition

Pruzansky's Technique	13 Bandpass Filter Parameters * 7 Bandpass Filter Parameters
Furui's Technique	*64 Cepstral Coefficients Derived from Average Spectra 12 Cepstral Coefficients Derived from Average Spectra
Sambur's Technique	LPC-10 Reflection Coefficients (Covariance Method) LPC-10 Filter Coefficients (Covariance Method) LPC-10 Log Area Ratios (Covariance Method) LPC-10 Reflection Coefficients (Autocorrelation Method) *LPC-12 Reflection Coefficients (Autocorrelation Method)
Markel's Technique	*LPC-10 Reflection Coefficients plus Pitch and Power (Covariance Method)

Speaker Recognition Feature Extraction

In the second functional block, the short term characterization of the speech (speech analysis parameters) is accumulated and speaker recognition features are produced. The speech analysis parameters are processed in two parts. The first part is the subpopulation filter which separates the input speech parameters into various classes. The subpopulations used for each technique in this study are given in Table 2.

Table 2: Subpopulation Classes Used for Speaker Recognition

Pruzansky's Technique	*All Speech Voiced Speech Unvoiced Speech Transitional Speech Vocalic Nuclei Nasals
Furui's Technique	*All Signal
Sambur's Technique	*All Speech
Markel's Technique	*Voiced Speech

The second part of the recognition feature extraction process is the accumulation of the mean vector and, in some cases, the covariance matrix, for the speech analysis parameters over some period of time. This process, which is referred to as "blocking", divides the data into blocks of 100, 200, 300, and 600 speech frames. Since each frame is 22.5 milliseconds long, this corresponds to approximately 2, 5, 7 and 13 seconds of speech respectively. The speaker recognition features produced are the first and second moments over a given time period of the speech analysis parameters in the various sub-population classes.

Model Generation

The third functional block is the model generator. In order to recognize a speaker, the system must store a model which characterizes each speaker of interest. The models are generated by characterizing the long term properties of the speaker using such parameters as the means and covariance matrix for the speaker recognition features over a substantial time interval.

Another factor in generating models is the portion of the data base actually used. For this study, two types of models were generated. The first type, time models, were generated from all six minutes of speech data from two sequential interviews. The second type, balanced models, used approximately one minute of speech from all ten interviews (ten minutes total). Figure 2 shows the portions of the data base used to generate two versions of each model type, as well as the portions used as unknown speech in each case.

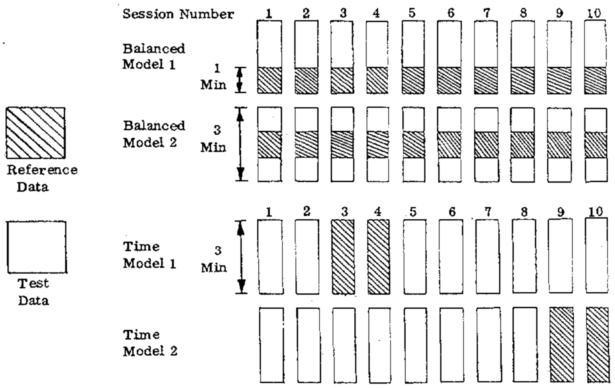

Figure 2. Data Base Subdivisions for Model Generation and Recognition Testing

Distance Metrics

The fourth functional block is the distance metric. In this block, the similarity between the stored speaker models and the unknown speaker recognition features is calculated. Most often the similarity is measured by calculating the "distance" between the unknown features and the features of the model. A decision as to the talker identity is then made on the basis of the model with the smallest distance to the unknown. All speaker distance functions used in this study can be written in the following form:

$$d_k^2 = (m - M_k)^t \; W_k \; (m - M_k)$$

where

d_k is the distance from the unknown feature m to model k, m is an n dimensional vector of the feature means from an unknown speaker,

M_k is an n dimensional vector of the feature means for speaker k,

W_k is an n x n matrix defined by the model type.

The distance metrics used for the four techniques are shown in Table 3. The main difference in the distance metrics are in the W matrix. The matrices used for each metric are shown in Table 4.

Table 3: Distance Metrics Used for Speaker Recognition.

Pruzansky's Technique	*Mahalanobis Metric Using Speaker Dependent Covariance Matrix. Weighted Euclidean Metric Using Speaker Independent Variances. Euclidean Metric.
Furui's Technique	*Mahalanobis Metric Using Both Speaker Dependent and Speaker Independent Covariance Matrices. Weighted Euclidean Metric Using Speaker Independent Variances. Euclidean Metric.
Sambur's Technique	*Mahalanobis Metric Using Speaker Dependent Covariance Matrix in an Orthogonalized Space.
Markel's Technique	*Mahalanobis Metric Using Speaker Dependent Covariance Matrix.

Table 4: W Matrix Used With Various Distance Metrics.

Mahalanobis Metric	W = Inverse Covariance Matrix from Model.
Weighted Euclidean Metric	W = Diagonal Matrix with Elements Equal to Inverse Variances.
Euclidean Metric	W = Identity Matrix.

EXPERIMENTAL DESIGN

In order to compare the four speaker recognition techniques fairly, a series of tests was designed to characterize the recognition techniques from a number of points of view. In all, over 300 separate recognition experiments were run. From the original four techniques, 23 separate recognition algorithms were formulated and evaluated. This multiplicity of algorithm variations resulted from the many combinations of speaker recognition features and speaker distance metrics in the four techniques. For example, with Sambur's technique, the recognition algorithm using orthogonal LPC is evaluated with three different LPC parameter sets and two different methods of determining speaker distances.

Each of the 23 recognition algorithms were evaluated using the same series of 14 recognition tests. This battery of 14 experiments was composed of two balanced models used with four recognition trial lengths (approximately 2, 5, 7, and 13 seconds) and two time models used with three recognition trial lengths (2, 5 and 7 seconds). In the testing of each technique, the test data used for each recognition trial was formed from the identical speech data across all techniques. Therefore, when evaluating the performance of the four techniques, it was possible not only to make valid comparisons on the overall performance, but also to make valid comparisons on a recognition trial by trial basis. To further aid in comparability, the models used for each of the 14 experiment types were developed using the same speech passages for all 23 recognition algorithms.

EXPERIMENTAL RESULTS

The results for the best algorithm for each of the four techniques are shown in Figure 3. Sambur's and Markel's techniques both performed well as text independent recognizers, achieving 95% and 93% accuracies with 10 minute balanced models and 13 second unknowns. These results are within a few percent of the original published results. The performance improves significantly as the recognition trial length is increased from 2 to 7 seconds of speech. Further increases in trial

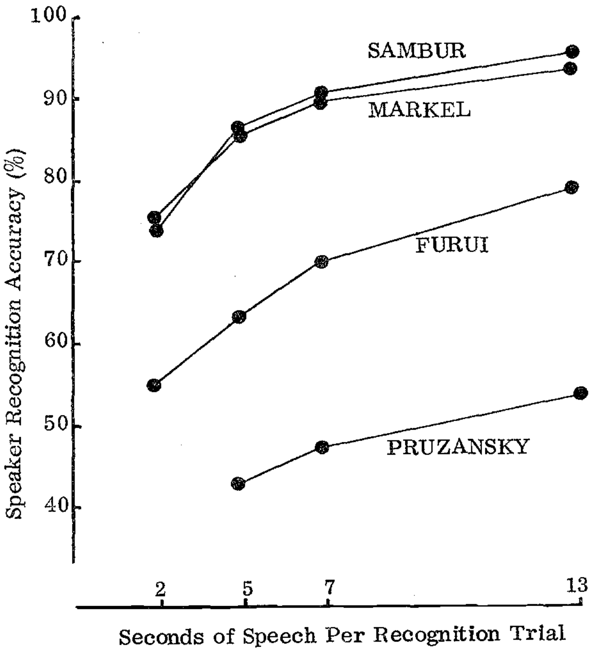

Figure 3. Comparison of Four Speaker Recognition Techniques with 10 Minute Balanced Models. Configurations are Indicated by Asterisks in Tables 1-3.

length appear to result in only slight performance improvements. However the percentage decrease in error rate continues to be almost linear with trial length.

The performance of Furui's technique is considerably less than that of Sambur's and Markel's. The overall low performance of this technique provided concern that the process was not implemented correctly. However, the technique was originally implemented in a text dependent environment, and tested with as much as 30 seconds of unknown speech. Although this study dealt with much shorter speech samples, the text independent implementation used in this study was verified by using a majority voting aproach to combine groups of three and five recognition trials into a single recognition. When the unknown speech used in each recognition trial was increased to 34 seconds, the recognition rate for Furui's technique rose sharply to 95%.

Pruzansky's technique achieved the lowest recognition rate. This technique was also originally implemented as a text dependent speaker recognizer and did not perform well in a text independent mode. One interesting result from Pruzansky's technique was the performance of the six subpopulations. The "all speech" population performed best, with the voiced, the unvoiced, and the transitional voiced speech subpopulations about 5% to 10% worse. The steady state nasals and the steady state vocalics performed poorly, probably due to the sparseness of these populations.

A result that was consistent across all techniques was the relative performance of the two model types. The balanced models (generated using speech from all interviews) always performed better than the time models (generated from two sequential interviews), and no significant differences were observed for the two versions within each model type. An unexpected result was that, although there was considerable variation in the recognition accuracies with the time models as a function of the interview session used as the unknown, there was no indication that performance decreased as the time between the model and the unknown recordings increased.

All of the models described to this point were generated with at least three minutes of speech. One question of interest was how sensitive was system performance to the amount of speech used in generating the models. Further testing of model generation was conducted using Markel's technique. Models were generated with 20 seconds of speech. Recognition accuracies of 94% were obtained for the 20 second models with 40 second recognition trials. These results, however, were for models generated from data recorded at the same session as the unknown. When models were generated using 20 seconds of speech recorded one week later than the unknowns, the recognition rate fell to 70%.

SUMMARY

This study investigated the performance of four different speaker recognition techniques in a text independent mode. It demonstrated that the two techniques which are based on LPC analysis (Sambur's and Markel's) perform well as speaker recognizers when large amounts (ten minutes) of speech are available for generating models. Markel's technique was also tested on models generated with as little as 20 seconds of speech, and was shown to maintain good recognition performance when the model data was recorded at the same session as the unknown.

The recognition systems based on spectral and cepstral parameters did not perform as well as the LPC based systems in the text independent mode. The superior performance of the LPC techniques is most likely due to the fact that LPC parameters are insensitive to changes in speech power and pitch. On the other hand, the parameters generated by the filter bank of Pruzanski's technique and the cepstral coefficients of Furui's technique, both derived from average spectra, are sensitive to the pitch and overall speech power. The large intra-speaker variations in these parameters encountered in the data base introduce significant changes in the long term average vectors used as recognition features, and consequently decrease system performance.

The performance of the time models indicates that there are significant changes in speaker characteristics from one interview session to another. However, over the ten week period spanned by the data base, there is no indication that the performance of the time models decreases as a function of time. The performance of the balanced models (ten minutes of speech recorded over a ten week period) indicates that models generated with speech from a number of differents sessions produce better estimates of speaker characteristics than can be obtained from a single session.

REFERENCES

1. Pruzansky S. and Mathews M.V., "Talker Recognition Procedure Based on Analysis of Variance," Journal Acoustical Soc Amer, Vol 36, No 11 pp. 2041-2047, Nov 1964
2. Furui S., Itakura F., and Saito S. "Talker Recognition by Longtime Averaged Speech Spectrum," Electronics and Communications in Japan, Vol 55A, No 10, pp. 54-62, Oct 1972
3. Sambur M.R. "Speaker Recognition Using Orthogonal Linear Prediction," IEEE Trans. on Acoustics, Speech and Signal Processing, Vol ASSP-24, No 4, pp. 283-289, Aug 1976.
4. Markel J.D., Oshika B.T., and Gray A.H. Jr., "Long-Term Feature Averaging for Speaker Recognition," IEEE Trans. on Acoustics, Speech, and Signal Processing,, Vol ASSP-25, No 4, pp. 330-337, Aug 1977.