A REALTIME IMPLEMENTATION OF A TEXT INDEPENDENT SPEAKER RECOGNITION SYSTEM

E. H. Wrench, Jr.

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

ABSTRACT

This paper describes the design and implementation of a realtime speaker recognition system. The system performs text independent, closed set speaker recognition with up to 30 talkers in realtime. In addition, the reference speech used to characterize the 30 talkers can be extracted from as little as 10 seconds of speech from each talker, and the actual recognition performed with less than one minute of speech from the unknown talker.

Two speaker recognition algorithms previously developed by Markel and Pfeifer were investigated for use in the realtime system. The results of this investigation clearly show that Markel's technique is superior for applications using very short speech segments for both the speaker models and the recognition trials. Markel's technique was implemented in realtime in a high speed programmable signal processor. A test of this implementation with a set of 30 male speakers resulted in recognition accuracies of 93-100% for models generated with only 10 seconds of speech, and recognition trials using only 10 seconds of unknown speech.

INTRODUCTION

The objective of this program was to demonstrate the feasibility of using speaker recognition techniques to aid in the identificatior of unknown speakers. The work was supported in part by Air Force contract F3062-78-C-0324 from the Rome Air Development Center. The most difficult task was to develop a text independent speaker recognition technique that would achieve high recognition accuracy with less than 30 seconds of reference data. Previous text independent speaker recognition techniques had been reported that achieved the required accuracies, but all had used several minutes of training data to generate the speaker reference models [1]. In addition, recognition was accomplished in most of these studies by using at least one minute of unknown speech. A serious question that needed to be investigated was whether or not the required accuracy could be maintained when the training data was reduced to 10 seconds, and the unknown speech was limited to less than one minute.

The program was divided into two parts. The first was to investigate speaker recognition techniques to determine their performance under the conditions of limited amounts of input speech. The second part was to implement the best technique in a laboratory demonstration system.

ALGORITHM SELECTION

Two speaker recognition techniques were investigated as to their performance using limited speech. The first was originally developed by Markel [2]. It was selected because of its performance in a previous speaker recognition study [1]. The second technique was originally developed by Pfeifer [3].

Markel's technique, as implemented for this study, is shown in Figure 1. It uses ten linear prediction coder (LPC) reflection coefficients from the voiced portions of the speech as speaker recognition features. The features are averaged over the entire recognition period, and the average feature vector is then compared with the stored talker models. The recognized talker is the one whose model is most similar to the unknown speech.

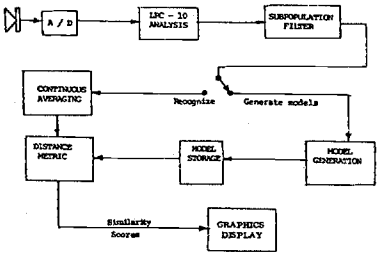

Figure 1: Markel's Speaker Recognition Technique

The model for each speaker is generated from the training data by calculating the mean vector and the covariance matrix for the reflection coefficients.

The distance metric used to determine the similarity between the unknown speech and each of the models is the Mahalanobis metric as defined in equation 1.

CH1610-5/81/0000-0193 $00.75

$$D = (F - M)^T [W]^{-1} (F - M) \qquad \text{Eq 1}$$

where F is the average coefficient vector,
M is the mean vector from a model,
and $[W]^{-1}$ is the inverse covariance matrix from the model.

Figure 2 shows Pfeifer's technique as implemented for this study. Pfeifer's technique also uses LPC reflection coefficients as speaker recognition features. The difference between the two techniques is that Pfeifer's algorithm does not average the features before comparing them with the stored models, but rather makes a decision as to the talker identity for every speech frame. The final recognition decision is then made by determining which model compared best with the unknown for the majority of the frames during the recognition period. Another difference between the two techniques is in the choice of subpopulation filters. Markel's technique used voice speech frames for both the models and the unknowns. Pfeifer's technique was originally implemented using steady state vowel-like speech for the recognition features. In this study, both steady state vowel regions and voiced speech were tried as recognition features.

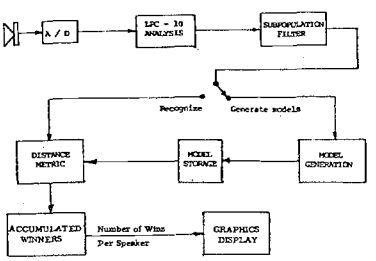

Figure 2: Pfeifer's Speaker Recognition Technique

The two techniques were evaluated using a data base consisting of 2 three minute interviews from 17 different talkers. The interviews with each talker were separated by a one week interval.

Experiments were conducted to investigate the performance of the two algorithms under a variety of conditions, all of which involved the use of limited amounts of input speech both for model generation and recognition. The majority of the testing used 10 or 20 seconds of speech for the reference models, and recognition used one to 40 seconds of speech.

The results for both Markel's and Pfeifer's techniques when used with 10 and 20 seconds of reference data are shown in Figure 3. The results indicate that Markel's algorithm performs better than Pfeifer's for applications where the durations of both the reference data and the model data are limited. The recognition performance of Markel's technique was 96% when used with 20 second models and 40 second unknowns. The results shown for Pfeifer's technique used voiced speech for recognition features. It was found that Pfeifer's algorithm performed better when voiced speech rather than steady state vowels were used as the recognition features. This is probably due to the fact that the speaker is not well characterized by the extremely small number of steady state vowel frames contained in 10 to 20 seconds of speech.

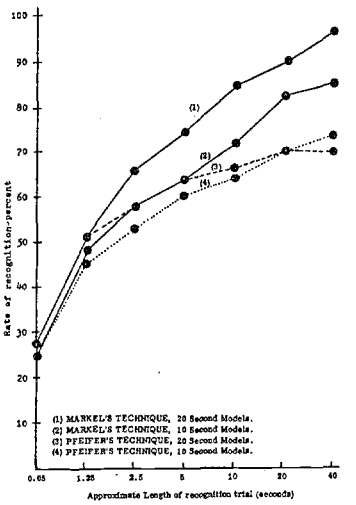

Figure 3: Comparison of Markel's and Pfeifer's Techniques

As discussed earlier, Markel's technique finds the model that most closely matches the average reflection coefficients from the unknown. Pfeifer's technique does no averaging before comparison with the models, but rather makes decisions frame by frame and uses a majority vote at the end of the unknown to identify the talker. The two techniques are not mutually exclusive. A combination of the techniques was implemented, where blocks of frames are averaged and a decision made for each block. A majority vote is then done at the end of utterance on the block by block decisions to identify the talker. Figure 4 shows a comparison of 6 experiments, all using 20 seconds of speech for each recognition trial. The performance is plotted as a function of the number of frames averaged to form a block divided by the number of blocks used in the majority vote. The product of frames per block times blocks per recognition is a constant 400 frames per recognition trial (~20seconds of speech). The results indicate that it is desirable to average as many frames as possible before calculating the distance to models, and not to make any intermediate hard decisions.

The performance of Markel's technique with noisy speech was evaluated for one particular signal to noise ratio. Figure 5 shows the performance of Markel's algorithm when the signal to noise ratio of the input speech was reduced to 15 dB. The models were generated using noisy training speech, and then the recognition trials were performed using noisy unknowns. As shown in the figure, the speaker recognition accuracy decreased by less than 10%.

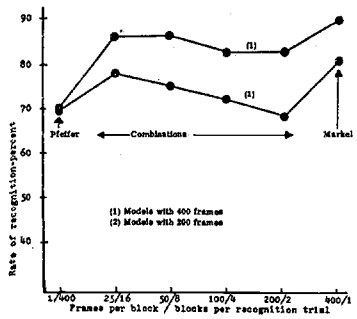

Figure 4: Combinations of Markel's and Pfeifer's Techniques
20 Second Recognition Trials

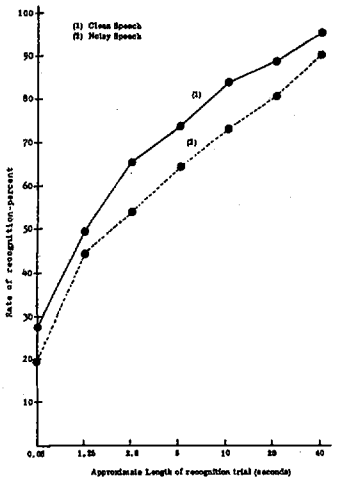

Figure 5: Performance of Markel's Technique With Noisy Speech
20 Second Models (15dB Signal to Noise Ratio)

Markel's algorithm was chosen for the laboratory realtime demonstration system based on the results obtained from the algorithm selection studies. The speaker recognition system is implemented using a PDP-11/60 as the system controller and an ITT developed, high speed signal processor, the Quintrell RAM, for the computationally intensive realtime processing. A block diagram of the system is shown in Figure 6. All operator interaction with the system is done through the PDP-11. The control program in the PDP-11 directs the Quintrell to perform the appropriate tasks. In addition, the PDP-11 is used for such functions as display formating, plotting and hard copy. The Quintrell is used for the processing that must be done in realtime, such as the LPC-10 analysis and the continuous averaging of coefficients.

One of the goals for the realtime system was to integrate it with an operator interface to provide a flexible, easy to use system. This was accomplished by prompting the operator at every phase of the operation with the options that are available at that time. The operator simply selects from a displayed menu the desired operating mode (model generation, recognition, etc.), and indicates the action to be taken via simple commands.

The system is capable of either model generation or speaker recognition for up to 30 talkers in realtime. During recognition, the system continuously displays the similarity between the unknown speaker and the stored models. A sample display is shown in Figure 7. The input speech from the unknown talker is analyzed frame by frame, and the LPC-10 reflection coefficients extracted. A voicing detection algorithm is used to discard all but the voiced frames. The voiced frames are then continuously averaged. A

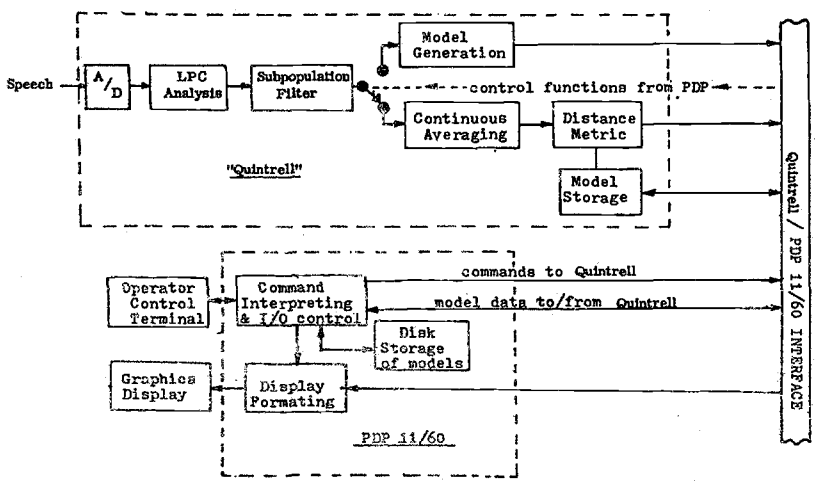

Figure 6: Laboratory Demonstration System Block Diagram

background process, which runs any time the LPC, voicing, or averaging routines are idle, calculates the Mahalanobis distance between the current average coefficient vector, and all of the models. With 30 models, the similarity between the unknown and every model is updated about once per second.

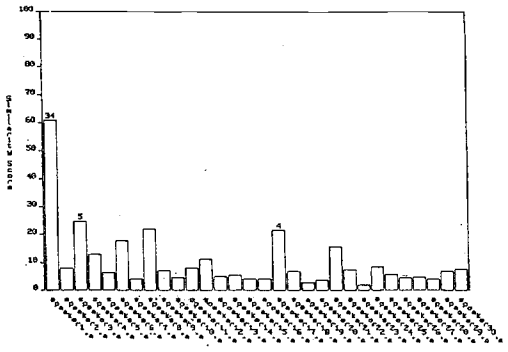

Figure 7: Sample Output From the Realtime Speaker Recognition System

Limited testing of the realtime speaker recognition system was conducted. A thirty talker data base was generated by recording speakers from commercial television. Speaker models were generated with 10 and 20 seconds of the recorded speech for each talker. Different 10 and 20 second segments from each talker were then used as unknowns.

The results are very encouraging. As table 1 indicates, the best recognition rate was 100% correct for 10 second models and 20 second unknowns. It is somewhat surprising that the 10 second models performed better than the 20 second models, however, it must be remembered that this was an extremely small test. Only one recognition trial was run for each speaker. Further testing is required to adequately estimate the system performance.

Table 1: REALTIME SPEAKER RECOGNITION RESULTS

Unknown	10 second models	20 second models
10 sec.	93% (28 correct out of 30)	90% (27 correct out of 30)
20 sec.	100% (30 correct out of 30)	97% (29 correct out of 30)

CONCLUSIONS

Markel's technique was shown to be well-suited for text independent speaker recognition with limited amounts of speech for both model generation and recognition. In addition, the study showed that Markel's technique performs better than the Pfeifer technique under these conditions.

The implementation of Markel's algorithm used in this study achieved recognition rates in excess of 90% for all speakers when used with limited amounts of speech for both the reference models and the unknowns. In addition, this algorithm has been implemented in a realtime speaker recognition demonstration system and achieves similar high recognition scores. The realtime demonstration system has proven to be easy to operate with little or no instruction. An operator can generate a model using "live" speech, document the model with pertinent speaker data, and use the model for realtime speaker recognition, all within less than a minute.

REFERENCES

1. Wohlford R.E., Wrench E.H., Landell B.P., "A Comparison of Four Techniques for Automatic Speaker Recognition" IEEE ICASSP 80 Proceedings, Denver, Colorado, vol 3, pp. 908-911. Apr 1980.

2. 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.

3. Pfeifer Larry L., "Feature Analysis for Speaker Identification" RADC-TR-77-277, Final Technical Report for Rome Air Development Center, August 1977.