AN APPROACH TO TEXT-INDEPENDENT SPEAKER RECOGNITION WITH SHORT UTTERANCES

K. P. Li
E. H. Wrench, Jr.

ITT Defense Communication Division
San Diego, CA 92131

ABSTRACT

A new technique for text-independent speaker recognition is proposed which uses a statistical model of the speaker's vector quantized speech. The technique retains text-independent properties while allowing considerably shorter test utterances than comparable speaker recognition systems. The frequently-occurring vectors or *characters* form a model of multiple points in the n dimensional speech space instead of the usual single point models. The speaker recognition depends on the statistical distribution of the distances between the speech frames from the unknown speaker and the closest points in the model. Models were generated with 100 seconds of conversational training speech for each of 11 male speakers. The system was able to identify 11 speakers with 96%, 87%, and 79% accuracy from sections of unknown speech of durations of 10, 5, and 3 seconds, respectively. Accurate recognition was also obtained even when there were variations in channels over which the training and testing data were obtained. A real-time demonstration system has been implemented including both training and recognition processes.

INTRODUCTION AND BACKGROUND

The two major approaches to speaker recognition are text-independent and text-dependent. Text-independence normally requires that long sections of speech be accumulated for both training and testing. Text-dependence requires a co-operative speaker and the locating and matching of like phonetic events in the model and the unknown speech. Text-dependent systems, however, achieve accurate recognition on very short speech segments. There are many applications in which only short, unconstrained speech segments are available for testing. The technique presented here is a melding of the text-independent and text-dependent approaches which allows speaker identification with unconstrained speech using test utterances as short as three seconds. Previous text-independent speaker recognition systems have utilized the longterm speech statistics to model and recognize speakers [1]. Since the systems must operate text-independently, relatively large amounts of speech are required for both the training and test utterances to determine the speaker statistics accurately. In most of these studies, at least 20 seconds of speech was required in the unknown utterance to achieve high accuracy recognition, particularly if the training and testing data are recorded at different times. The use of shorter duration test utterances usually causes severe degradation in performance due to the variability of the speech statistics. This variability is caused by differences in the phonetic content of the training and test data. Therefore, features other than longterm statistics need to be considered for text-independent speaker recognition if short utterances (less than ten seconds) are to be recognized.

In 1976, Wakita [2] reported an experiment on vowel recognition and speaker identification in which the distances between different vowels from different speakers were compared. He found that a speaker almost always produced vowels that were closer to one of his own vowel productions than to other speakers vowels, even if the matched vowels were not the same as the vowel being produced.

The results reported by Wakita suggest a new starting point for a speaker recognition algorithm. Any vowel uttered by a given speaker should have high probability of matching well with one of the vowels extracted from his training speech. Even in a text independent mode with unknown utterances of any length, all the vowels should match well with the reference set for the corresponding speaker.

A MODEL FOR TEXT-INDEPENDENT SPEAKER RECOGNITION

In the new technique, a speaker will be characterized as producing speech events that lie in several, potentially disjoint subspaces or *clusters* within the speech space. A reference model can be built up from a collection of points in the n dimensional speech space. These cluster centers approximate the multi-modal distribution of the speaker's characteristic sounds in the speech space, as shown in Figure 1. A set of speaker models will occupy overlapping regions of the space but will not be identical one with another. Unknown speech can be thought of as a sequence of points in the space. Successive measures of the distance of the unknown speech to each of the speaker models can be made. Our speaker recognition technique assembles the distances to the closest cluster for each unknown speech point. While an impostor's model may provide the smallest distance on any given unknown point, the lowest collection of speaker distances should come from the actual speaker's model.

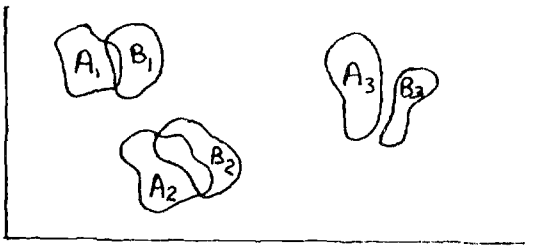

Figure 1: Clusters of Speech events for Speakers A and B in n Dimensional Speech Space.

The crucial problem in a technique such as this is to estimate the model, i.e., to locate the cluster center. For this, we will turn to the notion of character encoding. Character encoding is a means for covering the speech

ICASSP 83, BOSTON CH1841-6/83/0000-0555 $1.00 © 1983 IEEE

space with a set of irregular polyhedra where the centers of these polyhedra are generated from actual speech data. A typical covering is performed by setting a radius and inputting the spectral representation of a frame of speech. Typically, a frame is a 20 millisecond segment of speech. This frame forms the center of the first character. Subsequent frames whose distances from this or other characters exceed the radius are established as new centers of new characters. Frames which fall within the radius of a character are encoded as that character. In character encoding terms, then, the cluster centers for a speaker's model will be drawn from the set of characters used to encode that speaker.

In the ideal situation, there will be some subset of the characters used to encode a speaker's training data which are optimal for identifying the speaker during testing. The cluster centers for a speaker will be those points in the speech space for which the distances between these centers and the speaker's training data are smaller than the distances between the centers and any other speaker's training data. Thus, the characters to be selected for a particular speaker's model will be those that, on average, minimize the ratio of within speaker distances to impostor distances. Discovering this ideal set of characters is computationally expensive. In practice, we have made a heuristic approximation to the best set of cluster centers.

SELECTION OF SPEAKER REFERENCE MODELS

The steps in selecting speaker reference models were: First, the speech is encoded using LPC-12. Second, a threshold is set to test the overall amplitude of the sample to exclude those portions of the space occupied by silence or non-speech samples. Third, a simple covering technique is applied to generate a set of characters with a pre-determined radius, measured using the Itakura distance [3]. The number of characters generated in covering is dependent upon the size of the character radius. With a smaller radius, the number of characters is larger and the precision with which the speech events are described is increased. In our experiments, the size of the characters was set so that approximately 1000 characters covered the space used by all speakers. The 400 most frequently occurring characters, which cover more than 90% of total speech, were used to create the model.

For each speaker to be modeled, the following three steps are performed:

(1) the frequency of occurrence of each character is collected for a number of ten second speech samples.

(2) the mean and standard deviation for the frequency of occurrence of each character is estimated

(3) all other speakers are treated as impostors and their frequency of occurrence data pooled; the means and standard deviations for the impostors were then calculated for each character.

The list of speaker and impostor means and standard deviations for each character forms the data base necessary to construct the reference model.

There are two factors to consider in selecting speaker's references:

(1) the frequency of occurrence of a reference character must be higher for the speaker than for the impostors,

(2) the frequency of occurrence of a reference character for all the short duration samples must be stable; low standard deviation for both the speaker and the impostors.

There are many different combinations of these measures that could represent the effectiveness of a character in separating the speaker and impostors: F-ratio, t-ratio, or even a simple linear combination of the difference of means and sum of two standard deviations. We found that the application of several different measures only serves to change the ordering of the highest scoring 10% of the characters and does not affect which characters are chosen appreciably. Therefore, the recognition performance is not significantly affected by the choice of measure used to select characters for the model. All the experimental results presented in this study used a simple linear combination of means and standard deviations to select model characters.

We have selected a small number of characters, 40, for every speaker according to the criteria stated above. These selected characters for each speaker include about 10% of the total number of characters used, but cover from 25 to 40% of the total number of samples for each speaker. The model characters for a speaker cover only 7 to 12% of the impostor's samples. For the 11 speaker models used in our experiments, about 80% of most frequent 400 characters were used in the models, and about 20% of these characters were used in more than one speaker model. Thus, for a model consisting of 40 characters, selected as described, we can expect there to be a 25 to 40% chance for each short term frame of speech data having a close match with one of the 40 characters, whereas the chance for a frame from an impostor having a close match with the speaker's model is only 7 to 12%. This three to fourfold increase in probability forms the basis of the speaker recognition algorithm.

RECOGNITION PROCEDURES

In its most basic form, the recognition procedure is based on calculating the percentage of close matches between the unknown speech and each of the model character sets. For each model speaker, the frames from the test utterance are encoded with the characters contained in the speaker's model. The encoding process is accomplished by finding the closest model character to the unknown frame and saving the corresponding distance. These minimum distances are collected throughout the whole utterance to form a minimum distance distribution as shown in Figure 2a. This represents a statistical measure of the match between the unknown utterance and the model. For the model that corresponds to the test speaker, the minimum distances should be distributed at lower values than those for models of the impostors. The proportional higher frequency of occurrence of lower distances in the distribution indicates a high probability of match with the model. The unknown speaker is identified as the speaker corresponding to the model with the largest percentage of close matches.

Since the 40 selected model characters covered 25 to 40% of total training data, we may anticipate that when the model and the unknown are the same speaker, 25 to 40% of the minimum distances obtained in the character encoding process will be less than the radius used to generate the characters. For any impostor, only 7 to 12% of the minumum distances should be less than the character radius. The differences between these percentages are used to separate the speaker and the impostors.

A measure of the closeness of a match between an unknown utterance and a model is defined as that distance, D, that is greater than the minimum distance obtained from a fixed percentage of all the unknown frames. Figure 2b shows a typical distribution of minumum distance values for both a speaker and impostors. For the experiments reported here, the distance value, D, was set such that 30% of the minimum distance values for the unknown utterance would be less than D. If the unknown and the model match, the value of D should

be approximately the character radius, if they don't match, D will be larger than the character radius. This can be seen in Figure 2b.

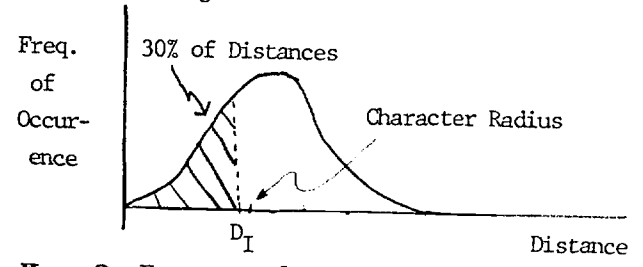

Figure 2a: Frequency of Occurance of Minimum Distances Between Speaker I and the Model for Speaker I.

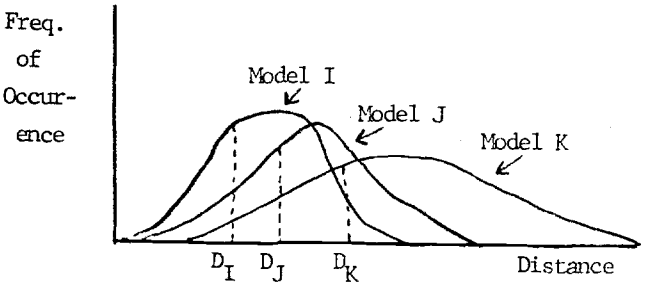

Figure 2b: Frequency of Occurance of Minimum Distances Between Speaker I and the Models for Speakers I, J, K.

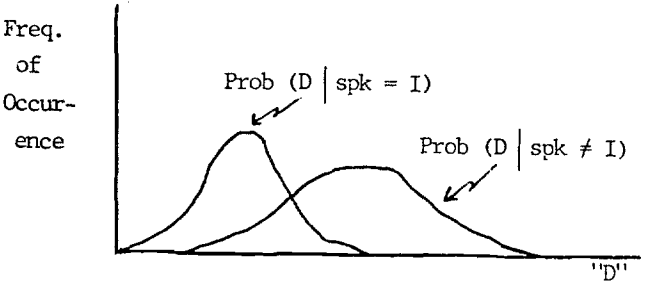

Figure 2c: Distribution of "D" Values using Speaker I's Model for Speaker I and Imposters.

During the model generation process, ten second segments of training data are used to determine a number of D values for both the correct speaker and the impostors for each model. A typical distribution of D values for a speaker and impostors is shown in Figure 2c. From these distributions of D's, the probability of a given D value given the correct speaker, $prob(D|speaker)$ and the probability of a given D value given an impostor, $prob(D|impostor)$ is estimated for all models. These probabilities become part of the speaker models, and are used in the recognition process. The stability of these distributions was checked by recalculating the distributions using test data, rather than training data. There are slight shifts in the means and changes in the standard deviation for speaker's distances, but the impostor's statistics remain stable.

During recognition, once a D value is determined from the unknown speech for a particular model, the likelihood ratio is determined for that model. The likelihood ratio is defined as:

$$L=\frac{prob(D|speaker)-prob(D|impostor)}{prob(D|speaker)+prob(D|impostor)}$$

An L value is calculated for each model, and the speaker corresponding to the model with the highest L is recognized.

EXPERIMENTAL RESULTS

The data base used for all the experiments reported here consisted of recordings of unconstrained coversational speech from 11 male speakers. The training data and testing data were recorded one week apart. The training data was 100 seconds of speech, including some silence, although all pauses were less than two seconds. Approximately 20% to 30% of speech samples are silence and not used for either modeling or recognition.

The recognition algorithm was tested in a variety of situations. The first test was to verify that the speaker models being generated were able to separate the speakers when the training data was used as the unknown speech. The result was 100% correct recognition, indicating that speaker characteristics were being modeled in the training data. The next experiment was to determine if the models could correctly identify speakers from independent test data recorded one week after the training data. The recognition was based on three different test utterance lengths, 10, 5, and 3 seconds. The recognition accuracy shown in Table 1 was 96%, 87%, and 79%, respectively.

Table 1: RECOGNITION RESULTS

Model Data	Unknown Data	Unknown Length		
		10 sec	5 sec	3 sec
Clean speech 10 sec/sample 10 sample/spk	Clean speech recorded one week after model	96%	87%	79%
Clean speech 3 sec/sample 33 sample/spk	Clean speech recorded one week after model	96%	---	80%

The models used for the experiments described above were generated with statistics estimated from ten training speech segments (ten seconds per segment) per speaker. To address the question of how the segment length affects recognition accuracy, particularly when the recognition trials are based on segment lengths less than ten seconds, new models were generated using statistics estimated from 33 segments each three seconds long. The results shown in Table 1 indicate that the length of the segments used have little effect on the recognition accuracy. This may indicate that the recognition results depend only on the total amount of training data, and the duration and number of segments used to estimate the statistics compensate each other.

In many practical applications, the test data may come from a different communication channel than the training data. One way to compensate for channel variation is to perform blind deconvolution to modify both the training and test data so that their longterm spectra are identical to some target spectrum. When this approach is applied to recognition techniques that use the longterm average of the speech parameters to model the speakers, the blind deconvolution tends to make all the speaker models look the same. The algorithm presented here does not have this problem, since it is the relative positions of the characters in the space that separates the speakers, not the position of one single character.

To test this hypothesis, several experiments were conducted. First, blind deconvolution was applied to both the training and test data to see if recognition results were severely affected by the blind deconvolution even if

the channels were the same. The experimental results shown in Table 2 indicate about 3% degradation for ten second duration unknowns. The second test was to verify that blind deconvolution could compensate for channel variations. The test data was modified by passing it through a telephone channel simulator to modify the longterm spectrum. The model data was left unmodified. Both the training data and the model data were then blind deconvolved to match a selected spectrum. The experimental results shown in Table 2 indicate that for ten second unknowns, the recognition accuracy is 8% less than when no channel differences were present. The third test was to perform blind deconvolution on each ten second unknown segment independently, rather than performing blind deconvolution on an entire unknown session as done above. The models were generated two ways. First, using data which had been blind deconvolved over the entire training session, and second, using data which had been blind deconvolved in ten second segments, the same as for the test data. Table 2 show a significant degradation in performance when the blind deconvolution is applied differently for the training and testing data. However, if blind deconvolution is independently applied to each ten second segment in both the test and the training data, the recognition results are about the same as that when one blind deconvolution is applied to the entire test and training session. This is a major improvement over the results obtained with other recognition techniques where channel normalization severely degrades recognition accuracy.

Table 2: BLIND DECONVOLUTION EXPERIMENT
Ten Second Recognition Trials
Test Data 1 Week after Model Data

Model Data	Test Data	% Correct
Clean Speech 100 seconds	Clean speech	96%
Blind Deconvolution Clean Speech	Blind Deconvolution Clean Speech	93%
100 Second Blind Deconvolution Clean Speech	Blind Deconvolution Channel Distorted Speech	88%
100 Second Blind Deconvolution Clean Speech	Channel Distorted Speech Blind Deconvolution every 10 seconds	77%
10 Second Clean Speech Blind Deconvolution every 10 seconds	Channel Distorted Speech Blind Deconvolution every 10 seconds	89%

The final experiment involved the use of the speaker probabilities to set rejection thresholds for the recognition system. If the probabilities (likelihood ratios) for the two most probable speakers are within 5% of each other, then the system rejects the sample and does not make a decision. A comparison of the results with and without rejections is shown in Table 3. The recognition rates are significantly improved with a reasonable rejection rate. For the case with blind deconvolution of channel distorted data, the recognition rate increased from 89% correct with 0% rejection to 95% with about 11% rejection.

REAL-TIME IMPLEMENTATION WITH ARRAY PROCESSOR

The algorithm described here has been implemented in realtime on a VAX-11/780 computer and a Floating Point Systems AP-120B array processor. The input speech is digitized and stored on disk. As the digitizing proceeds, the data is read from the disk, and transferred to the array processor. Once in the array processor, LPC-12 analysis is performed, and the data is character encoded using model characters stored in the data memory of the array processor. The minimum distances are saved in local storage until an entire utterance has been processed. At that time, the data is sent back to the VAX where the likelihood ratios are calculated and recognition results are output. The algorithm currently runs in real-time with 8K sample-per-second data using 12.5 millisecond frames.

Table 3: RECOGNITION RESULTS WITH REJECTIONS
Ten Second Unknown Speech Samples
Test Data 1 Week after Model Data

Data	% Correct	% Rejected
Clean Speech	96% 97%	0% 7.2%
Channel Distortion Speech Blind Deconvolved Every 10 Seconds	89% 95%	0% 11%

CONCLUSION

A new approach to speaker recognition has been described which is capable of recognizing a speaker from very short utterances. A statistical model is used to represent a multi-modal distribution of the speaker's vocal characteristics. The model can be used for either speaker verification or speaker identification. Recognition rates in excess of 95% have been achieved for ten second unknown speech samples from 11 male talkers, even when the test and training data were recorded during different weeks. The experimental results presented indicate that accurate recognition is obtained even when there are variations in channels over which the training and testing data are obtained. A real-time demonstration system has been implemented on VAX-11/780 with an FPS array processor.

The major differences between this technique and other text-independent speaker recognition systems are:

(1) The speaker model is multi-modal, rather than uni-modal.

(2) The decision is based only on the close matches with the model, rather than all the unknown speech. Events in the unknown not observed in the training data will not affect the recognition significantly.

(3) The decision is made over the set of individual matches, rather than the match of a single long-term average frame.

ACKNOWLEDGEMENT

This research was supported by the Air Force Systems Command, Rome Air Development Center, Griffiss Air Force Base. The authors also express their appreciation to T. Ward, S. Johnson, and P. Berens for programming support, to J. Naylor for performing the channel simulation and blind deconvolution experiments, R. Wohlford for numerous discussions and encouragement, and to P. Benson for reworking the Manuscript.

REFERENCES

1. J. D. Markel , B. T. Oshika , and A. H. Gray Jr. , "Long-term feature averaging for speaker recognition," *IEEE Trans. on Acoustics, Speech, and Signal Processing* **ASSP-25** pp. 330-337 (1977).
2. H. Wakita , "Residual energy of linear prediction applied to vowel and Speaker Recognition," *IEEE Trans. on Acoustics, Speech, and Signal Processing* **ASSP-24** pp. 270-271 (1976).
3. F. Itakura , "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. on Acoustics, Speech, and Signal Processing* **ASSP-23** pp. 67-72 (1975).