SPEAKER VERIFICATION USING TEMPORAL DECORRELATION POST-PROCESSING

Lorin P. Netsch and George R. Doddington†

Texas Instruments Incorporated
Computer Science Laboratory
PO Box 655474, MS 238
Dallas, Texas 75265

ABSTRACT

This paper describes a text-dependent method of speaker verification processing which utilizes the statistical correlation between measured features of speech across whole words. The correlation is used in a linear discriminant analysis to define uncorrelated word-level features as a metric. Initial results indicate that this method can significantly reduce the amount of storage necessary for speaker specific speech information. Further, this method provides promise of improved verification performance compared to methods based on HMM state level observation metrics. Since the linear discriminant analysis yields features which are decorrelated over entire words, this method should be more robust to signal distortions which are consistent over the entire utterance.

1. INTRODUCTION

Identity verification is a requirement of many applications. Verification by voice is attractive since it involves a natural response of the user. Existing algorithms [1]provide acceptable performance in restricted environments where the microphone, background noise, and speech may be controlled. However, many speaker verification applications must rely on available distorted speech which degrades speaker verification performance. A typical example is telecommunications, where use of different telephone handsets and, channels cause the distortion. Many recent studies [2] [3] [4]address this distortion problem.

This paper describes an enhancement to previously reported algorithms [4] that may be viewed as a post-processing step occurring after a recognition task that utilizes HMM-based word models. Since only a single handset or channel is used during a call, it seems plausible that the mechanism causing the distortion is consistent temporally. In this case, it should be possible to find statistical features that span entire words and which are relatively independent of the distortions encountered. Further, one suspects that there is significant correlation between observed features corresponding to the states from a known HMM word model. Thus, rather than calculating an utterance likelihood based on the individual statistics of each state of the HMM word models, a method is constructed which calculates the likelihood based on statistics of entire words.

The verification system implemented uses a vocabulary consisting of the digits "zero" through "nine" and "oh". It is assumed that each valid speaker utters a known digit string determined at enrollment.

2. DESCRIPTION

Verification system front-end processing which creates word-level features is shown in Figure 1 and described in the following paragraphs.

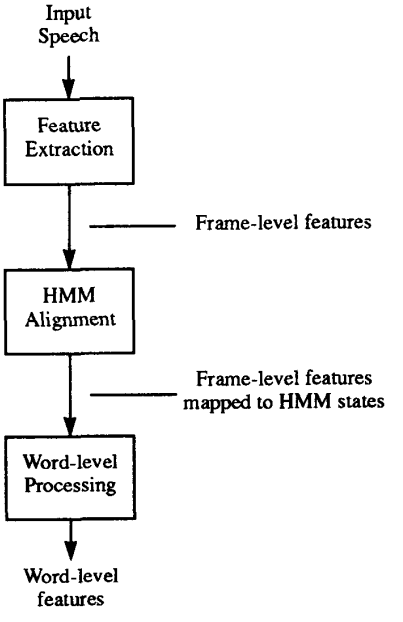

Figure 1. Verification Processing

† Presently with SRI International, Menlo Park, Ca.

0-7803-0532-9/92 $3.00 © 1992 IEEE

2.1. Feature Extraction

Speech entering the system is sampled at 8kHz and processed to produce a stream of speech feature vectors. The features are created from a tenth order autocorrelation LPC model calculated every 20 msec using a 30 msec Hamming window. The LPC parameters are used to create a spectral energy vector of a synthetic filter bank with 12 mel-spaced filters. Each speech frame is represented by a 27-element speech parameter vector which is composed of both current-frame parameters and short-term (40 msec) difference parameters:

1) The overall rate of short-term spectral transition integrated over frequency (in dB)
2) The frame energy, relative to long-term averaged speech level (in dB)
3-14) The filter energy, relative to the frame energy (in dB)
15) The short-term change in frame energy (in dB)
16-27) The short-term change in filter energy (in dB)

The 27 element speech parameter vector is then transformed to a 16-element feature vector by a linear transformation which is designed to normalize covariance statistics of the feature vector [5].

2.2. HMM Alignment

2.2.1. Alignment

Alignment of input speech feature vectors to HMM word model state observations is accomplished by a conventional speaker-independent digit recognition algorithm. The speech is aligned according to the Viterbi maximum likelihood path through word model HMMs with continuous multivariate Gaussian state observation models. Since the claimed identity corresponds to a known digit string, this string is used (along with the possibility of inter-word non-speech) as a grammar to constrain the recognition. The result of the recognition is a mapping of input speech feature vectors to digit HMM word-model state observations.

2.2.2. HMM models

The speaker-independent digit HMM word model topology designed for this verification algorithm allows for variation in speech rate and also ensures that at least one input speech feature frame will be aligned with each state of the model. The number of states in the HMM word model for each digit varied proportionally to the nominal duration of the spoken digit. Multiple states shared the same observation model, and the number of observation models was selected to be half of the nominal duration in frames. Separate HMM models were created for male and female speakers. Figure 2 shows a typical HMM word model. The vertical states in the model all share the same observation models. Beginning and ending observation models are restricted to explain a finite duration of speech. This was done to limit the effect of leading and trailing non-speech on the observation model parameters.

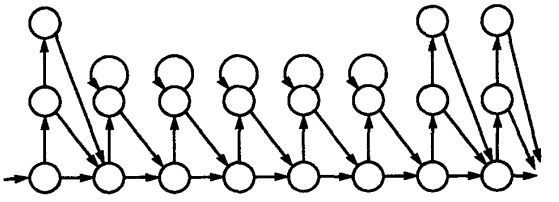

Figure 2. HMM word model for "four"

2.3. Word-level processing

2.3.1. Word-level vectors

Post-recognition word-level processing constructs a single vector for each word of the input utterance. Each word-level vector is formed by first averaging all input feature vectors mapped to each HMM observation model to form an averaged input utterance feature vector for each observation. We then concatenate the averaged feature vectors corresponding to each word to form the word-level vector. Note that each word-level vector has a dimension which is the product of the number of frame level features (16) and the number of observation models in the HMM for the word.

2.3.2. Word-level features

The word-level vectors corresponding to each of the words undergo a temporal decorrelating linear transformation (TDT) unique to each digit to form a 20-element word-level feature vector. The transformation is designed to produce decorrelated word-level features for each word which are used to calculate the likelihood. The transformation normalizes the covariance of the within-speaker statistics and also provides a linear discriminant measure between speakers.

2.3.3. Calculating the TDT

The TDT for each digit is calculated using speech data in which many speakers provide many tokens of the digit. We first calculate a covariance matrix and mean of word-level vectors (described in 2.3.1. above) corresponding to the digit for each speaker. Pooling the covariance matrices for each speaker results in a single "pooled within-speaker" covariance matrix. The covariance matrices and means corresponding to the digit for all speakers are then used to calculate a "total" covariance matrix which represents the covariance of all word-level vectors from all speakers. A linear discriminant analysis [6] is performed

using the "pooled within-speaker" and "total" covariances, and the 20 features which best discriminate between speakers are retained to form the TDT. Both male and female data was used to form the TDT for each digit.

3. Experimental Evaluation

3.1. Speech corpora used

Two speech corpora were used in the evaluation of the verification method presented.

3.1.1. Verification 1 corpus

This corpus consists of 25 female and 25 male subjects. Each subject provided 25 calls over long-distance telephone lines, and was encouraged to use a wide variety of handsets and environments. During the call each subject provided three tokens of a unique familiar ten-digit string (a leading digit plus a social security number), and three tokens of the unique digit string of another subject of the same gender.

3.1.2. Verification 2 corpus

This corpus consists of 36 female and 70 male subjects. Each subject provided 25 or more calls over long-distance telephone lines, and was encouraged to use a wide variety of handsets and environments. During the call, each subject provided three tokens consisting of a four-digit string and two other non-digit words.

3.2. Experimental paradigm

The test speakers consisted of the last twelve male and twelve female speakers in the Verification 1 corpus.

3.2.1. Speaker enrollment

Enrollment was accomplished using the first call for each of the test speakers. Hence, there were three tokens of each test speaker's ten-digit phrase used for enrollment. Enrollment consisted of aligning the test speaker's input speech vectors with the speaker-independent HMM models, and calculating the average word-level feature vector for each word. The averaged word-level feature vectors were stored as the reference vectors for the speaker.

3.2.2. True speaker verification scoring

After enrollment, true speaker verification scoring utilized the remaining 24 calls for each test subject. For each verification token, we calculated word-level feature vectors for each word, and formed a word score as the Euclidean distance between the input word-level feature vector and the stored reference word-level feature vector. The final verification token score is the sum of each word score weighted by the relative duration of the input word.

3.2.3. True speaker reference updating

The reference word-model feature vectors were updated using the last token of each true speaker call. Updating consisted simply of averaging the input word-level feature vectors with the reference word-level feature vectors.

3.2.4. Impostor verification scoring

Impostor verification scores were calculated after all true speaker verification scoring and reference updating. Scoring for the impostors was performed in a manner similar to true speaker scoring. Impostors were always the same gender as the true speaker.

3.3. Experimental results

3.3.1. Verification 1 experiments

This set of experiments used the Verification 1 corpus only to form the TDTs. The TDTs were formed using data from the first 13 male and 13 female subjects. A single TDT was calculated for each digit using data from both males and females. After forming the TDTs, two experiments were performed. A closed set experiment used the first 13 male and 13 female subjects to calculate true speaker and impostor scores. An open set experiment used the last 12 male and 12 female subjects. The resulting ROC for each experiment is shown in Figure 3. Also shown for reference is our best performing closed set ROC based on calculation of likelihood using each individual HMM observation, modeled as a multivariate Gaussian distribution.

3.3.2. Augmented Covariance

A second set of experiments used the same test subjects as in 3.3.1., but utilized the Verification 2 corpus digit

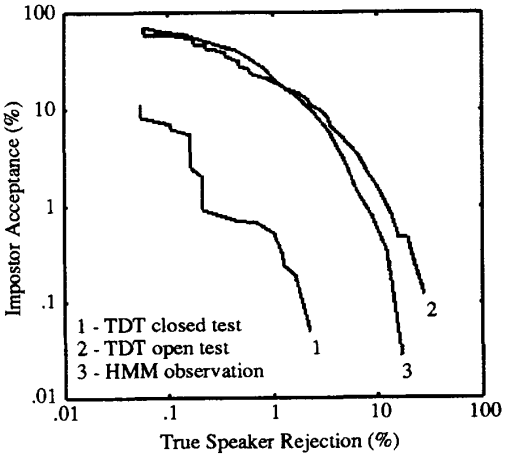

Figure 3. Total-voice ROC

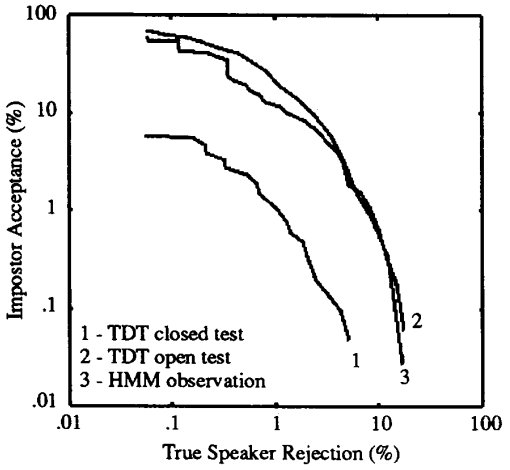

Figure 4. Augmented Covariance ROC

data to augment the "within-speaker covariance" and the "total" covariance estimate in calculation of the TDTs. The resulting ROC for each experiment is shown in Figure 4. Also shown again for reference is same ROC based on individual HMM observation models.

4. DISCUSSION

From the initial ROC results of Figure 3, it can be seen that the word-level feature verification algorithm for the closed set test provided improved performance over our closed set HMM observation model approach. In addition, the speaker-specific information needed per word has been reduced significantly. For an HMM observation model metric algorithm, one must store a 16-element feature vector corresponding to each observation model in the HMM (typically this results in 128 values for our HMM structure). Our experiments used only 20 word-level features, which provides a significant reduction in speaker-dependent storage needed. The word-level feature performance for the open set test in Figure 3, however, did not provide improved performance over the HMM observation model approach.

Of importance in the results is the discrepancy between the closed-set and open-set results. The superior performance of the closed-set results in Figure 3 implies that the TDT's are matched to the closed-set subjects used to generate the TDTs. This may have implications for speaker identification algorithms. When data from the Verification 2 corpus were used to augment word-level covariance estimates in Figure 4, as expected, closed-set performance degraded while open-set performance improved. This indicates the problem in calculating the TDTs, which is that estimation of the large dimension of word-level covariance matrices requires large amounts of data. We are presently involved in further collection of digit data over long-distance telephone lines that will improve covariance estimates. One can at least be assured that the best open-set performance lies somewhere between the closed-set and open-set results in Figure 4.

REFERENCES

[1] J.M. Naik and G.R. Doddington, "High Performance Speaker Verification Using Principal Spectral Components", *Proceedings of ICASSP '86*, Vol. 2, April 1986, pp. 881-884.

[2] T. Matsui and S. Furui, "A Text-Independent Speaker Recognition Method Robust against Utterance Variations", *Proceedings of ICASSP '91*, Vol. 1, May 1991, pp. 377-380.

[3] R.C. Rose, J. Fitzmaurice, E.M. Hofstetter, and D.A. Reynolds, "Robust Speaker Identification in Noisy Environments Using Noise Adaptive Speaker Models", *Proceedings of ICASSP '91*, Vol. 1, May 1991, pp. 401-404.

[4] J.M. Naik, L.P. Netsch, and G.R. Doddington, "Speaker Verification Over Long Distance Telephone Lines", *Proceedings of ICASSP '89*, Vol. 1, May 1989, pp. 524-527.

[5] G.R. Doddington, "Phonetically Sensitive Discriminants for Improved Speech Recognition", *Proceedings of ICASSP '89*, Vol. 1, May 1989, pp. 556-559.

[6] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. Orlando, Fl.: Academic, 1972.