TEXT-DEPENDENT SPEAKER VERIFICATION USING DATA FUSION

Kevin R. Farrell

Dictaphone Corporation
3191 Broadbridge Avenue
Stratford, Connecticut, USA 06497
farrekr@pb.com

ABSTRACT

A new system is presented for text-dependent speaker verification. The system uses data fusion concepts to combine the results of distortion-based and discriminant-based classifiers. Hence, both intraspeaker and interspeaker information are utilized in the final decision. The distortion and discriminant-based classifiers are based on dynamic time warping (DTW) and the neural tree network (NTN), respectively. The system is evaluated with several hundred two word utterances collected over a telephone channel. The combined classifier yields an equal error rate of two percent for this task, which is better than the individual performance of either classifier.

1. INTRODUCTION

The objective of speaker verification is to verify a person's claimed identity based on an utterance from that person. This is in contrast to speaker identification where a person is to be identified within a population. Speaker verification is generally considered to be more commercially applicable than speaker identification.

A distinguishing feature of speaker verification systems regards the form of spoken input, which can be text dependent or text independent. Text-dependent speaker recognition systems require that the speaker utter a specific phrase or a given password. Text-independent speaker identification systems identify the speaker regardless of the utterance. Text-independent systems are more convenient from a user standpoint in that there is no need for a password. However, in order to completely model and evaluate the acoustic feature space of a speaker, substantially more data is necessary for training and testing than for text-dependent systems. This paper focuses on text-dependent speaker verification.

Early work on text-dependent speaker verification utilized dynamic time warping (DTW) [1]. Later, it was found that hidden Markov models (HMMs) yielded improved performance over DTW techniques [2]. Several forms of HMMs including subword models [3] and whole word models [4], have since been considered. In general, HMMs have been considered the predominant technology for text-dependent speaker verification. However, one limitation of HMMs is that they generally require more data than other classifiers to sufficiently learn the model parameters. This is a serious limitation for many commercial applications where it is desirable for the enrollment and authentication to be brief.

A new system is proposed for text-dependent speaker verification. The new system uses data fusion principles to combine the results of two sets of text-dependent speaker models. The first set of speaker models utilizes neural tree networks (NTNs) [5]. NTNs are discriminant-based and provide an interspeaker measure. The NTNs have been successfully applied to text-independent speaker recognition [6] where, specifically, advantages were found for speaker verification along with tasks that involved limited training data. The second set of text-dependent speaker models uses a DTW-based technique. The DTW method uses a distortion measurement between the extracted features of a test utterance and those stored for that speaker during training. Hence it provides an intraspeaker measure. Additionally, it uses temporal information, which currently is not used in the NTN. Since the information provided by the NTN and DTW classifiers is somewhat independent, they can be combined to yield a powerful classification system for text-dependent speaker verification.

This paper is organized as follows. Section 2 provides an overview of specific task considered in this paper in addition to a description of the NTN and DTW methods for creating text-dependent speaker models. Section 3 describes the methods for combining the results of the NTN and DTW models. Section 4 provides experimental results and the summary and conclusions of the paper are given in Section 5.

2. CLASSIFIERS

The task considered in this research focuses on an application that requires authenticating speakers over a telephone channel based on limited training and testing data. Specifically, four utterances of a password are used for training and the verification decision is based on one utterance of that password. This constraint confines the classification approaches that can be used in the system. The approach considered here is described as follows.

For the four training utterances, not only must a model be trained, but additionally an appropriate threshold must be determined. A resampling technique, namely the "leave-one-out" method is used here to accommodate the small number of training exemplars. Three utterances are used to

This work was done while the author was with the CAIP Center at Rutgers University and SpeakEZ.

0-7803-2431-5/95 $4.00

train a speaker model and the remaining utterance is used as an independent test case. This procedure is repeated four times, each time leaving out a different test utterance, to yield four models.

The two speaker modeling methods considered in this paper are based on the neural tree network (NTN) and dynamic time warping (DTW). The NTN bases its decision upon discriminant information whereas the DTW method utilizes a distortion measure. Hence, these two methods use criteria that is somewhat complimentary. The NTN and DTW methods of speaker verification are briefly described as follows.

2.1. Neural Tree Network

The NTN [5] is a hierarchical classifier that combines the properties of decision trees and feed-forward neural networks. For speaker recognition, the training data for the NTN consists of data for the target speaker labeled as "one" and data from other speakers labeled as "zero". The NTN partitions feature space into regions that are assigned probabilities which reflect how likely a speaker is to have generated a feature vector that falls within that region. The NTN has been evaluated for text-independent speaker recognition [6] where it was found to perform favorably for open-set problems, such as speaker verification.

There are several approaches for applying the NTN to *text-dependent* speaker recognition. One method is to train the NTN to discriminate between a password spoken by the target speaker and the same password spoken by other speakers. This can be interpreted as a "whole-word" NTN. Another method is to build "sub-word" NTN models [7]. In this case, the password of the target speaker will be segmented into sub-words, i.e., phonemes. A NTN will then be trained for each sub-word unit, where the anti-class data will consist of the data from other speakers' utterances of that sub-word unit. These models can then be concatenated to form passwords. The "whole-word" NTN will be considered here. Note that this method does not utilize the temporal information of the password. This information can be obtained, however, by combining the results of the NTN with a temporal-based model, such as that obtained from the DTW method.

2.2. Dynamic Time Warping

The DTW algorithm is a distortion-based approach for time aligning the dynamics of two waveforms. For speaker verification, a reference template can be generated from several utterances of the password [1]. These utterances are combined to form a single reference template using a modified k-means algorithm [8]. Then during testing, a decision can be made to accept or reject the claimed identity based on whether or not the distortion falls below a predetermined threshold. To allow for subsequent fusion with other speaker models, such as NTNs, the DTW distortions must be converted to a compatible scale, i.e., a probability. We accomplish this by simply raising the negative distortion to an exponential.

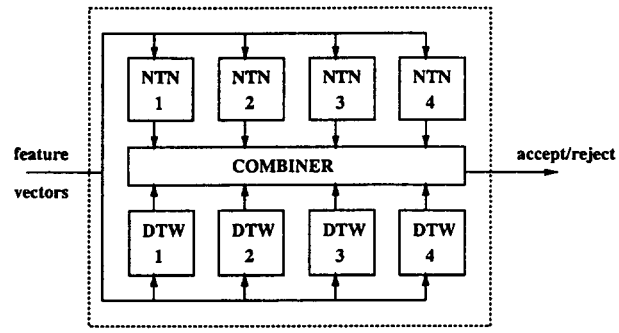

Figure 1: Verification system for claimed identity

3. COMBINED CLASSIFIER SYSTEM

The combined classifier system uses the outputs of both sets of classifiers for its final decision. Each individual classifier will provide its decision and a confidence to the combiner, which will output the final decision. This concept is illustrated in Figure 1.

The decision of each individual classifier is determined from the score as compared to a threshold. The threshold here is computed as a function of the *intraspeaker* and *interspeaker* scores. The intraspeaker score for each classifier is found as the performance on the left out utterance. Similarly, the interspeaker score is found by applying imposter utterances to the model and computing the average of the top scores (nominally five).

Each of the eight classifiers has its own threshold which is determined as follows:

$$T = x * interspeaker + y * intraspeaker, \quad (1)$$

where T is the threshold. Typical values for x and y are 0.8 and 0.2, respectively. During testing, each classifier will output a one or zero, corresponding to whether or not the score is greater than the threshold. Additionally, the score of each classifier will be output.

There are numerous methods for combining this information. One method is to simply take a vote on the different classifiers [9], i.e., if at least five classifiers verify the claimed identity, then accept it. Other methods utilize weighted sums or products of the classifier outputs, which are known as *linear opinion pools* and *log opinion pools* [10]. These two methods are evaluated here. The linear opinion pool consists of a weighted sum of the classifier outputs:

$$P_{linear}(x) = \sum_{i=1}^{n} \alpha_i p_i(x), \quad (2)$$

where $P_{linear}(x)$ is the probability output by the combined system, α_i are weights, $p_i(x)$ is the probability output by the i^{th} classifier, and n is the number of classifiers. For all experiments in this paper, α is between zero and one and the sum of the α's is equal to one. The linear opinion pool has been considered in speaker recognition for the combination of features [11], namely cepstrum and delta cepstrum features, in addition to combining NTNs and vector quantization classifiers for text-independent speaker recognition [12].

The log opinion pool consists of a weighted product of the classifier outputs:

$$P_{log}(x) = \prod_{i=1}^{n} p_i^{\alpha_i}(x). \quad (3)$$

Several approaches are considered here for combining the scores of speaker models. The first approach is the voting method. This method requires the final decision of each model, which is determined by whether or not the score exceeds a threshold. The voting method can be viewed as a model that provides an integer speaker score ranging from zero to eight. In essence, this quantizes the speaker score. The alternatives to the voting method that are considered here are the linear and log opinion pools. These methods utilize the average score for each set of models, i.e., the average NTN and average DTW score. Hence, the linear and log opinion pool methods described in equations (2) and (3) will be used to combine two scores and not eight.

4. EXPERIMENTS

The database used to evaluate the system was collected over a telephone channel. All of the handsets utilized electret microphones. The speech acquired through the telephone channel is sampled at 8 kHz, and μ-law coded at 8 bits/sample. The speech signal is pre-emphasized with a pre-emphasis factor of 0.95. Features are extracted within 20 millisecond analysis windows having 5 millisecond shifts between consecutive analysis windows. The features extracted from the analysis windows consist of linear prediction (LP)-derived cepstral coefficients.

The system is evaluated with data from 20 male speakers. Fourteen utterances of the state, "New Jersey" are collected from ten of these 20 speakers. Four utterances are used to train each model and the remaining ten are used for testing. For the remaining ten speakers, ten utterances of the state, "New Jersey" are collected and used as imposter utterances for each of the first ten speakers. This database allows for 100 true speaker trials and 1000 imposter trials. Note that a pre-existing database is used to provide the antispeaker data for the NTN training. This antispeaker database contains four utterances of "New Jersey" from 20 male speakers. The 10 imposter speakers used to evaluate the false accept rate are not included in the antispeaker database.

The operating curves for the NTN and DTW were evaluated and are shown in Figure 2. It is noted that these curves are based on a posterior positioning of the threshold. Here it is seen that the NTN performs significantly better than DTW, i.e., an equal error rate of 5.2% as compared to 9.8%. However, the equal error rate of the combined system using the linear opinion pool with $\alpha = 0.5$ is 2%, which is better than that of either method used individually.

The error rates of the linear and log opinion pools are evaluated as a function of α. The performance of both methods is illustrated in Figure 3. The best equal error rates for each method, including the voting method described in the previous section, are provided in Table 1. Here, it can be seen that all forms of data fusion yield a reduction in the equal error rate to that of either method

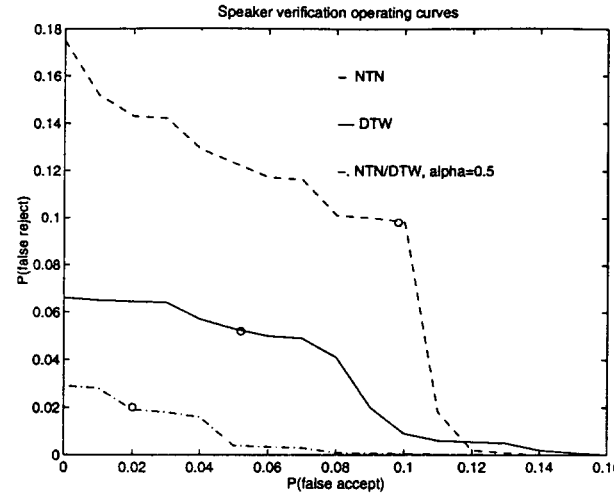

Figure 2: Speaker verification operating curves

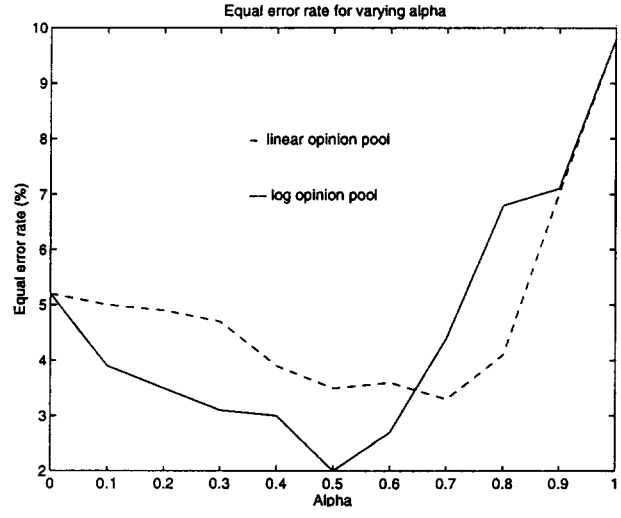

Figure 3: Equal error rate versus alpha

used individually. The linear opinion pool is found to provide the best overall performance.

An additional aspect that has been relatively unexplored for using speaker models based on supervised training algorithms is the amount of antispeakers that should be used in the training set. The results shown in Figures 2 and 3 use the data from all 20 speakers in the antispeaker training set. An experiment was performed to evaluate the sensitivity of the NTN to the number of antispeakers in the training set. The results of this experiment are shown in Figure 4. Here, the equal error rate (EER) is evaluated for NTNs trained with one antispeaker, then two antispeakers, etc. Four random orders of antispeaker selections are evaluated and the average EER for each antispeaker number is plotted. Here, it can be seen that for antispeaker populations of six and greater, the EER varies between roughly five and six percent. This result shows that a NTN can achieve good performance with as little as six antispeakers.

Table 1: Equal Error Rates

	NTN	DTW	linear	log	vote
EER	5.2%	9.8%	2.0%	3.3%	4.0%

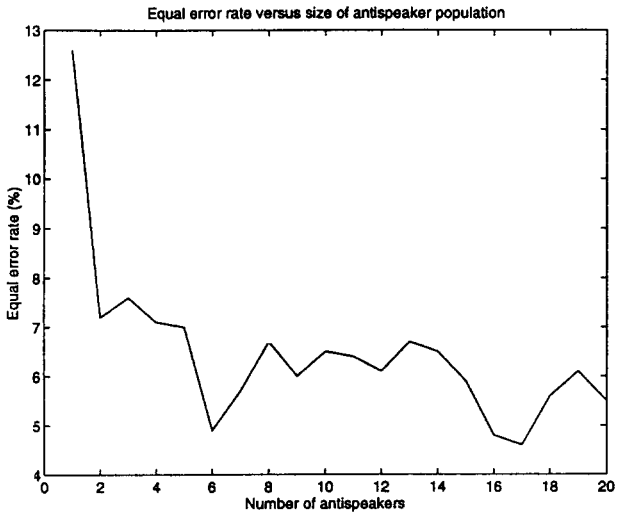

Figure 4: EER vs. number antispeakers

5. CONCLUSION

A new system is presented and evaluated for text-dependent speaker verification. The system uses data fusion concepts to combine the results of NTN and DTW speaker models. The system is evaluated with several hundred short utterances collected over a telephone channel. Several methods are considered for combining the classifiers, namely voting and the linear and log opinion pools. The equal error rate for the voting method is 4%. The equal error rates for the optimal linear and log opinion pools are 2% and 3.3%, respectively. All combined classifier approaches surpass the individual performance of the classifiers, which is 5.2% and 9.8% for the NTN and DTW methods, respectively.

6. REFERENCES

[1] S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, ASSP-29:254–272, April 1981.

[2] J.M. Naik, L.P. Netsch, and G.R. Doddington. Speaker verification over long distance telephone lines. In *Proceedings ICASSP*, 1989.

[3] A.E. Rosenberg, C.H. Lee, and F.K. Soong. Sub-word unit talker verification using hidden Markov models. In *Proceedings ICASSP*, pages 269–272, 1990.

[4] A.E. Rosenberg, C.H. Lee, and S. Gokeen. Connected word talker recognition using whole word hidden Markov models. In *Proceedings ICASSP*, pages 381–384, 1991.

[5] A. Sankar and R.J. Mammone. Growing and pruning neural tree networks. *IEEE Trans. on Computers*, C-42:221–229, March 1993.

[6] K.R. Farrell, R.J. Mammone, and K.T. Assaleh. Speaker recognition using neural networks and conventional classifiers. *IEEE Trans. Speech and Audio Processing*, 2(1), part 2, 1994.

[7] H. Liou and R.J. Mammone. Text-dependent speaker verification using sub-word neural tree networks. In *SPIE Proceedings, Conference on the Automatic Inspection and Identification of Humans*, July 1994.

[8] J.G. Wilpon and L.R. Rabiner. A modified k-means clustering algorithm for use in isolated word recognition. *IEEE Trans. ASSP*, 33:587–594, June 1985.

[9] L. Xu, A. Krzyzak, and C.Y. Suen. Methods of combining multiple classifiers and their applications to handwritten character recognition. *IEEE Trans. on Systems, Man and Cybernetics*, 23(3):418–435, 1992.

[10] J.A. Benediktsson and P.H. Swain. Consensus theoretic classification methods. *IEEE Trans. on Systems, Man and Cybernetics*, 22(4):688–704, 1992.

[11] F. K. Soong and A. E. Rosenberg. On the use of instantaneous and transitional spectral information in speaker recognition. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, ASSP-36:871–879, June 1988.

[12] K.R. Farrell and R.J. Mammone. Hybrid vector quantization/neural tree network classifiers for speaker recognition. In *SPIE Proceedings, Conference on the Automatic Inspection and Identification of Humans*, July 1994.