AN ANALYSIS OF DATA FUSION METHODS FOR SPEAKER VERIFICATION

Kevin R. Farrell[1] *Ravi P. Ramachandran*[2] *Richard J. Mammone*[3]

[1]T-NETIX/SpeakEZ Inc., 67 Inverness Drive East, Englewood, Colorado 80112
[2]Rowan University, Glassboro, New Jersey 08028
[3]CAIP Center, Rutgers University, Piscataway, NJ 08855

ABSTRACT

In this paper, we analyze the diversity of information as provided by several modeling approaches for speaker verification. This information is used to facilitate the fusion of the individual results into an overall result that provides advantages in accuracy over the individual models. The modeling methods that are evaluated consist of the neural tree network (NTN), Gaussian mixture model (GMM), hidden Markov model (HMM), and dynamic time warping (DTW). With the exception of DTW, all methods utilize subword-based approaches. The phrase-level scores for each modeling approach are used for combination. Several data fusion methods are evaluated for combining the model results, including the linear and log opinion pool approaches along with voting. The results of the above analysis have been integrated into a system that has been tested with several databases collected within landline and cellular environments. We have found the linear and log opinion pool methods to consistently reduce the error rate from that obtained when the models are used individually.

1. INTRODUCTION

Speaker verification consists of determining whether or not a voice sample provides sufficient match to a claimed identity. Speaker verification is a problem within the field of pattern recognition where it is desired to distinguish the identity of a person from that of other persons based on a voice sample. Speaker verification systems are either text-dependent or text-independent. Text-dependent systems require that the same password be used for both training and testing, whereas text-independent systems do not constrain the speech used for testing to be the same as that used for training. This paper provides analysis for text-dependent speaker verification.

Text-dependent speaker verification systems typically use modeling approaches that incorporate temporal information within the model. Traditional modeling approaches to text-dependent speaker verification include hidden Markov models (HMMs) [1, 2, 3] and dynamic time warping (DTW) [4] techniques. Neural network approaches, including whole-word neural tree networks (NTNs) [5] and subword NTNs [6] have also been evaluated for text-dependent speaker verification.

In this paper, we analyze the diversity of information provided by several popular modeling approaches for text-dependent speaker verification. Four modeling approaches are considered for score combination, namely the neural tree network (NTN), Gaussian mixture model (GMM), hidden Markov model (HMM), and dynamic time warping (DTW). Several data fusion methods are evaluated for combining the scores of the four models. These consist of the linear and log opinion pool methods along with voting.

This paper is organized as follows. The following section provides a description of the modeling approaches that we consider in this study. Then several methods for combining the outputs of the different modeling approaches are provided. This is followed by a description of the implementation details that are specific to our system. The experimental results for several text-dependent tasks are then provided. The databases used for these experiments are collected within both landline and cellular environments. A summary of the results is then given.

2. SPEAKER VERIFICATION MODELING

The models that are evaluated in this paper consist of the neural tree network (NTN), Gaussian mixture model (GMM), hidden Markov model (HMM), and dynamic time warping (DTW). These models have all been used in isolation for speaker verification. The GMM and HMM are both based on parametric representations of the feature space occupied by the speaker. The main difference here is that the HMM incorporates transitional information whereas the GMM does not. The DTW method is based on a distortion measure that is computed between a template representing an average of the enrollment utterances and a given test utterance. The NTN provides a discriminative-based speaker score that is based on a measure which incorporates information from other speakers. These modeling approaches are discussed below in more detail.

2.1. Neural Tree Network

The NTN [7] is a hierarchical classifier that uses a tree architecture to implement a sequential linear decision strategy. Specifically, the training data for a NTN consists of data from a target speaker, labeled as one, along with data from other speakers that are labeled as zero. The NTN will then learn to distinguish regions of feature space that belong to the target speaker from those that are more likely to belong to an impostor. These regions of feature space will correspond to leaves in the NTN that contain probabilities of the target speaker having generated data that falls within that region of feature space [8].

For sub-word NTN implementations, there is a NTN trained for each segment of data. The segmented data can be obtained by first evaluating the utterance with a hidden Markov model (HMM). The data for each segment is then evaluated with the corresponding NTN and the leaf probabilities for all observations are averaged across the phrase.

The NTN has been evaluated for text-independent speaker verification [8], whole-word based, text-dependent

0-7803-4428-6/98 $10.00

speaker verification [5], and subword based, text-dependent speaker verification [9, 6]. Data fusion methods were considered for whole-word NTN models with dynamic time warping [5, 10].

2.2. Gaussian Mixture Model

The Gaussian mixture model (GMM) has been evaluated for numerous tasks within speaker recognition [11, 12]. Essentially, a region of feature space for a target speaker is represented by a multivariate Gaussian distribution. The test vectors for an unknown speaker can then be evaluated with the distribution of the target speaker to determine the score. The score for the GMM as used in this paper is computed as

$$\hat{p}(X|S=S_t) = \frac{1}{N}\sum_{i=1}^{N} e^{-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)}. \quad (1)$$

where μ and Σ are the mean vector and diagonal covariance matrix, respectively, for the current segment. The expression in equation(1) is similar to a multivariate Gaussian distribution. However, we have removed the normalization term $\frac{1}{2\pi \det \Sigma^{-1}}$.

2.3. Hidden Markov Model

The continuous hidden Markov model (HMM) considers state transition information in addition to the mixtures that are modeled by the GMM. The HMM has been evaluated for speaker verification using both whole-word [2] and subword [3] models. The HMM has also been found to perform favorably to DTW [1] for certain speaker verification applications. The HMM is currently one of the most popular modeling approaches for text-dependent speaker verification systems.

2.4. 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 [4]. Then during testing, a decision can be made to accept or reject the claimed identity based on whether or not the distortion measured to the training template falls below a predetermined threshold. To allow for subsequent fusion with other speaker models, the DTW distortions must be converted to a compatible scale, i.e., from zero to one. We accomplish this by simply raising the scaled negative distortion to an exponential.

3. DATA FUSION

In this paper, we evaluate several methods for combining the output scores from the NTN, GMM, HMM, and DTW models. These consist of the linear opinion pool, log opinion pool, and voting. These methods are now described in more detail.

The linear opinion pool method computes the final score as a weighted sum of the outputs for each model:

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

where $P_{linear}(x)$ is the probability of the combined system, α_i are weights, $p_i(x)$ is the probability output by the i^{th} model, and n is the number of models.

Password text	# speakers devel/eval	# trials true/impostor
"Open Sesame"	47/56	195/11,229
Own full name	80/10 males	100/450
"Al Capone"	15/26	273/6825

Table 1. Specifications for the databases

Another approach for combining data is the log opinion pool. The log opinion pool consists of a weighted product of the classifier outputs:

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

The linear and log opinion pool methods are both simple approaches for combining the outputs of different modeling approaches. Though the performance from both techniques tends to be comparable, it has been noted that the output distribution for the log opinion pool method must be unimodal whereas this is not necessarily the case for the linear opinion pool method [13]. This can lead to a simpler decision strategy for the log opinion pool method.

Voting [14] is also evaluated here to combine the output decisions of the different models. The output of the vote is a score between zero and four where four refers to the case in which the score exceeds the threshold for all models. The final model decision is based on how many votes an utterance receives. The success of the voting method relies heavily on the selection of threshold for the individual models.

4. EXPERIMENTAL RESULTS

The modeling approaches and data fusion methods are evaluated with three toll quality speech corpora that were collected in house. The first database is known as the "full name" database. The full name database consists of 10 enrolled male target speakers and 80 development speakers. Each enrolled speaker has three enrollment utterances of their full name. The imposter attempts are obtained from the remaining nine speakers and all use the correct password. The second database is known as the "open sesame" database. This database consists of 56 enrolled speakers and 47 development speakers. Each speaker enrolled with the phrase "open sesame", hence, this scenario reflects a fixed-text situation. The third database is known as the "cellular" database. This database is also a fixed-text application that uses the password "Al Capone" for all speakers. This database was collected using cellular phones and consists of 26 enrolled speakers and and 15 development speakers. The aspects of each database are summarized in Table 1. The develop speakers column in Table 1 refer to the *development* set that is used to train a NTN. To avoid bias in the results, the development speakers are not used as imposters during the actual testing. The *evaluation* speakers are used to measure the actual system performance.

For the full name database, an analysis is performed to determine the correlation between errors for the respective models. In order to evaluate the models with the same performance criteria, we set the thresholds for each model to have zero percent false reject. Hence, all errors are false accept errors. The error correlation between models is shown in Table 2.

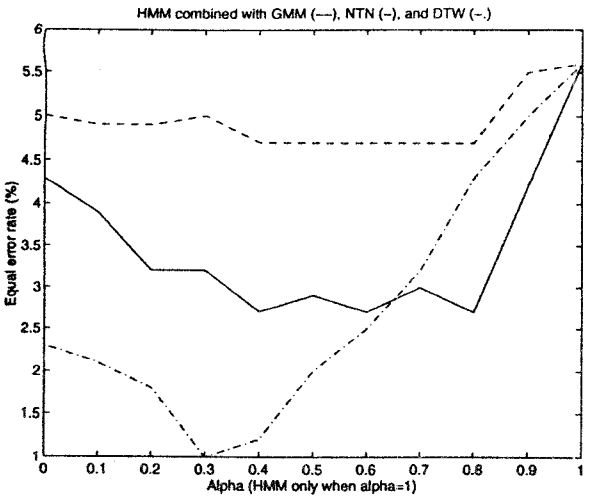

Figure 1. Data fusion results for HMM

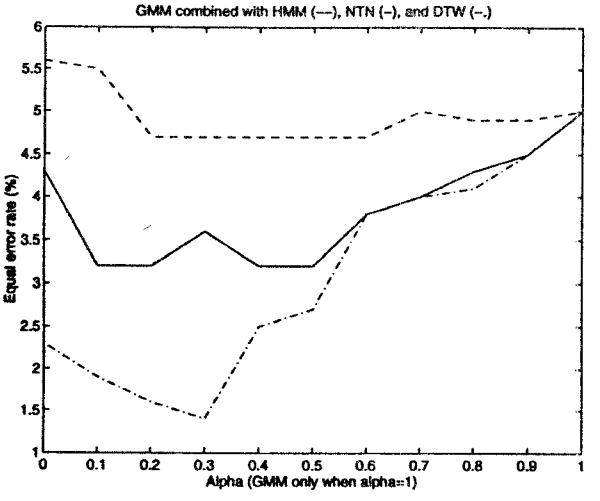

Figure 2. Data fusion results for GMM

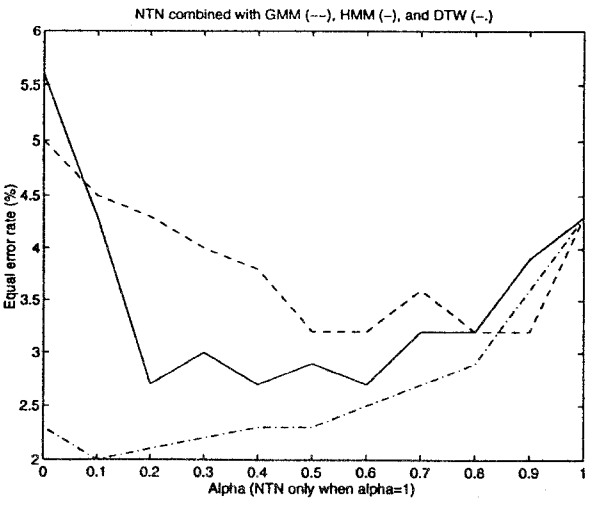

Figure 3. Data fusion results for NTN

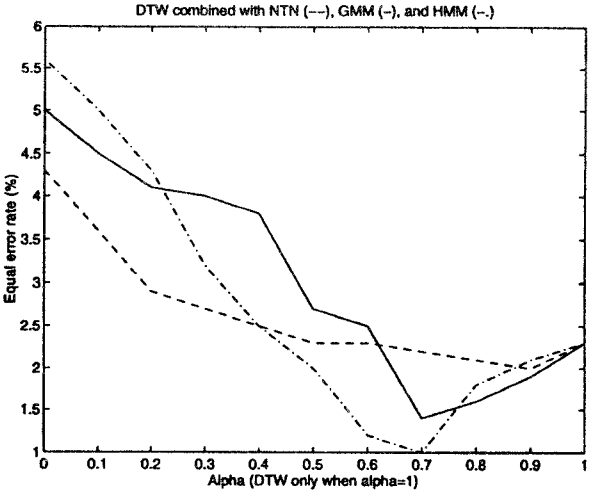

Figure 4. Data fusion results for DTW

The diagonal entries in Table 2 correspond to the false accept error of the model for the case of zero false rejects. In order for data fusion to be successful, models should be selected for combination that have a minimal correlation between errors. For example, if two models are to both make an error on the same observation then no combination of their scores will rectify the result. With this in mind, one would then expect the best HMM performance to obtained by combining the HMM with DTW and the worst performance by combining the HMM with the GMM. This is indeed the case as is illustrated in Figure 1 where the linear opinion pool is used for combining the data.

The same evaluation is provided for the GMM, NTN, and DTW models as shown in Figures 2 through 4, respectively. In virtually all of the cases it is shown that the best and worst model pairs can be determined by the correlation between errors.

The best equal error rates that are obtainable using the linear opinion pool approach are listed in Table 3. From Table 3 it can be seen that the two best performing models for this database are the DTW and NTN approaches. However, the best performance for model pair is obtained by combining the HMM with DTW. This result is consistent with the information provided in Table 2 where it can be seen that the errors between the HMM and DTW are less correlated than those between the NTN and DTW.

Model	HMM	GMM	NTN	DTW
HMM	0.085	0.056	0.035	0.029
GMM	0.056	0.084	0.049	0.036
NTN	0.035	0.049	0.075	0.038
DTW	0.029	0.036	0.038	0.055

Table 2. Error correlation between models

Model	HMM	GMM	NTN	DTW
HMM	5.6	4.7	2.7	1.0
GMM	4.7	5.0	3.2	1.4
NTN	2.7	3.2	4.3	2.0
DTW	1.0	1.4	2.0	2.3

Table 3. Best equal error rates between models

Model	Full Name	"Open Sesame"	"Al Capone"
NTN	4.3 %	1.8 %	7.1 %
DTW	2.3 %	4.3 %	7.1 %
GMM	5.0 %	3.5 %	9.6 %
HMM	5.6 %	2.8 %	9.6 %
Vote	2.3 %	3.1 %	8.2 %
Linear	2.1 %	2.0 %	7.0 %
Log	2.1 %	1.8 %	6.7 %

Table 4. EER performance for fusion methods

The results of the individual models in addition to those obtained from the linear opinion pool, log opinion pool, and voting methods for combining all four models are shown in Table 4. The weights for the linear and log opinion pool methods are based on Fisher discriminant analysis [15] where the inter and intra scatter matrices are computed from the pooled speaker and imposter data. By comparing the results for the full name database in Tables 4 and 3, it can be seen that the results in Table 4 are not optimal. This is due to the pooled scatter matrices being used as opposed to computing this information separately for each individual speaker. In Table 4, the linear and log opinion pool methods generally yield better performance than that obtained from the models used individually. This performance can be improved by adjusting the model weights on a speaker by speaker basis as opposed to using the same weights across all speakers. The voting method did not perform as well due to the sensitivity of threshold selection for the individual models.

5. CONCLUSION

Data fusion methods are evaluated for sub-word, text-dependent speaker verification. The different modeling approaches that are evaluated in this study include the hidden Markov model (HMM), Gaussian mixture model (GMM), neural tree network (NTN), and dynamic time warping (DTW). An analysis of the correlation between errors for these four modeling approaches is provided. The analysis shows that the best performance after combination is that obtained by combining the models with the least correlation between errors as opposed to combining the models that have the best performance. Data fusion methods are then evaluated for combining the results of all models. The data fusion methods include the linear opinion pool, log opinion pool, and voting. These methods are evaluated for several databases that include data collected within both cellular and landline environments. The linear and log opinion pool methods performed the best overall when evaluated for these tasks.

REFERENCES

[1] J.M. Naik, L.P. Netsch, and G.R. Doddington. Speaker verification over long distance telephone lines. In *Proceedings ICASSP*, pages 524–527, 1989.

[2] 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.

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

[5] K.R. Farrell. Text-dependent speaker verification using data fusion. In *Proceedings ICASSP*, 1995.

[6] M. Sharma and R.J. Mammone. Subword-based text-dependent speaker verification system with user selectable passwords. In *Proceedings ICASSP*, 1996.

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

[8] 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.

[9] H. Liou and R.J. Mammone. Text-dependent speaker verification using sub-word neural tree networks. In *Proceedings ICASSP*, 1995.

[10] K.R. Farrell. Discriminatory measures for speaker recognition. In *Proceedings of Neural Networks for Signal Processing*, 1995.

[11] H. Gish, M. Schmidt, and A. Mielke. A robust, segmental method for text independent speaker identification. In *Proceedings ICASSP*, pages 145–148, 1994.

[12] D. Reynolds. Speaker identification and verification using Gaussian mixture models. *Speech Communications*, 17:91–108, August 1995.

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

[14] 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.

[15] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis.* Wiley, New York, NY, 1973.