SUB-WORD SPEAKER VERIFICATION USING DATA FUSION METHODS

Kevin R. Farrell, Ravi P. Ramachandran, Manish Sharma, and Richard J. Mammone

T-NETIX/SpeakEZ Inc.
67 Inverness Drive East
Englewood, Colorado 80112
Tel: 303-705-5556, FAX: 303-790-9540
email: kevin.farrell@denver.t-netix.com

ABSTRACT

peaker verification is a rapidly maturing technology that is beoming available for commercial applications. In this paper, we ıvestigate the application of data fusion methods to sub-word imlementations of speaker verification. At a sub-word level, we utize the diversity of the information provided by the neural tree etwork and Gaussian mixture model to provide a more robust ub-word model. The phrase-level scores for each modeling aproach are obtained and then combined. The data fusion method 'e use for combining the model scores is the linear opinion pool. In ddition to using the diversity of the model scores, we also apply he concept of redundancy by using a leave-one-out approach to artition the input data. This allows us to generate several models nd accommodate the small training sample issues imposed by our ɔecific applications. The theoretical results of the above analy-s have been integrated into a system that has been tested with ɜveral databases that were collected within landline and cellular ıvironments. These results are included in this paper. We have ɔund that the proper data fusion techniques will typically reduce ıe error rate by a factor of two.

3-4256-9/97/$10.00 ©1997 IEEE

1 INTRODUCTION

Speaker verification consists of determining whether or not a voice sample provides sufficient match to a claimed identity. Speaker verification has numerous applications in areas that necessitate the validation of a person's identity. For example, when initiating a bank account transaction over the phone or at an automatic teller machine (ATM), speaker verification can provide an additional level of security over personal identification numbers (PINs). Also, speaker verification has the advantage over other forms of biometric authentification, such as fingerprint, retinal scan, etc., in that it can be applied over the telephone network. These are some of the characteristics that make speaker verification a very attractive technology for numerous commercial applications.

Speaker verification applications are generally text-independent or text-dependent. Text-independent speaker verification systems do not require that the same text be used for training and testing. Text-dependent speaker verification systems require that the same text be used during both training and testing. Though text-independent systems may be more convenient from a user standpoint, text-dependent systems provide additional security in that they 1) require fraudulent imposter attempts to use the same password, and 2) tend to provide better performance than text-independent systems. Text-dependent speaker verification systems will be the focus of this paper.

In this paper, we investigate the application of data fusion methods to sub-word model implementations of text-dependent speaker verification. The effects of segmentation for sub-word implementations are addressed. Two modeling approaches are then considered for score combination, namely the neural tree network and Gaussian mixture model.

This paper is organized as follows. The following section provides an overview of the processing steps in performing speaker verification. This overview includes a brief description of feature extraction, model evaluation and data fusion. 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

Speaker verification generally consists of feature extraction followed by model construction and evaluation. As part of model construction and evaluation we will also address the concept of data fusion where the scores of several models are combined to create a composite score. This composite score will be that which is applied to a threshold to yield the final decision. These phases of speaker verification are briefly described in the following subsections.

.1 Feature Extraction

eature extraction consists of deriving characteristics of the speech signal ıat are unique to an individual. The predominant characteristic that causes eople's voices to be different from one another is the shape of the vocal tract. he difference in the length and cross-sectional areas in the vocal tract from erson to person results in different resonant frequencies and bandwidths. ence, most feature extraction routines for speaker recognition utilize some pe of spectral analysis. Typical features are the cepstrum or variants of . Pole-filtered, mean-removed cepstrum [1] are the features used in the perimental results section. For this feature set we first obtain a channel timate by computing the pole-filtered mean of the linear predictive (LP) epstrum of the input speech. This channel estimate is converted to a filter ıat is applied to the speech to inverse out the channel effect. Then, the LP epstrum of the filtered speech is used as the feature.

.2 Modeling

speaker verification model is constructed from feature data, specifically ıat from a target speaker and possibly from non-target speakers. This model ıould have the ability to provide a level of match to the target speaker when ven a new set of feature data. For text-dependent speaker verification, a ıodel should capture the temporal information in addition to the acoustical formation. The standard models that accomplish this are hidden Markov odels (HMMs) and dynamic time warping (DTW). In general, segment-ısed approaches to speaker verification maintain temporal information. An-ıer important piece of information for model construction or evaluation is ıta that is not from the target speaker, or "non-target" data. One method r incorporating this information is used during model evaluation and is ıown as cohort normalization [2]. Another method is to use non-target data ıring training, which can be accomplished by using discriminative training proaches [3] or neural networks [4].

The modeling approach here is based on the neural tree network (NTN) ıd Gaussian mixture model (GMM). The NTN [5] is a hierarchical classifier at uses a tree architecture to implement a sequential linear decision strat-y. The NTN has been evaluated for text-independent speaker verification , whole-word based, text-dependent speaker verification [6], and sub-word ısed, text-dependent speaker verification [7, 8]. Data fusion methods were nsidered for whole-word NTN models with dynamic time warping [6, 9]. this paper, we evaluate data fusion methods for sub-word NTN models mbined with Gaussian mixture modeling, which is also a popular model for eaker verification [10].

2.3 Data Fusion

Data fusion methods can take advantage of the concepts of diversity and redundancy to improve system performance. Diversity can be used to improve system performance through the incorporation of different information. Similarly, redundancy can achieve the same goals through the re-use of data. These concepts have been thoroughly explored in the field of communications and have also been applied to pattern recognition problems. The basic idea is that if several models can be constructed, whose errors are mutually uncorrelated, then performance advantages can be obtained through the proper combination of the model scores.

The combination of different sources of information has been explored within a field known as data fusion. A comparison was done between several data fusion techniques, including the linear and log opinion pools [11], and voting [12] for a speaker verification application [6]. This comparison showed the simplest method, namely the linear opinion pool, to do at least as well as the other methods. Hence, the linear opinion pool will be considered here. The linear opinion pool is evaluated as a weighted sum of the outputs for each model:

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

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. For all experiments in this paper, α_i is between zero and one and the sum of the α_i's is equal to one.

3 SPEAKER VERIFICATION SYSTEM

The speaker verification system used in this paper is known as the T-NETI *SpeakEZ Voice Print*^{SM} system. This system is text dependent and utilizes sub-word NTN and GMM models, along with vocabulary-independent password selection and data fusion. The vocabulary-independent password selection is enabled through a technique known as blind segmentation [13]. The blind segmentation algorithm will automatically determine the number of segments and segment boundaries for a password without the use of transcription information. The NTN and GMM scores for each subword are accumulated to form the phrase-level score for each model type.

Additionally, a leave-one-out strategy is deployed to utilize the data redundancy in addition to facilitating threshold selection. Basically, for enrollment repetitions of a password, there will be N separate models. Each model is trained with $N-1$ repetitions with a different repetition "left-out" for each model. The left-out repetition can then be applied to the model to yield an unbiased target speaker score that can be used in setting the threshold for speaker acceptance/rejection.

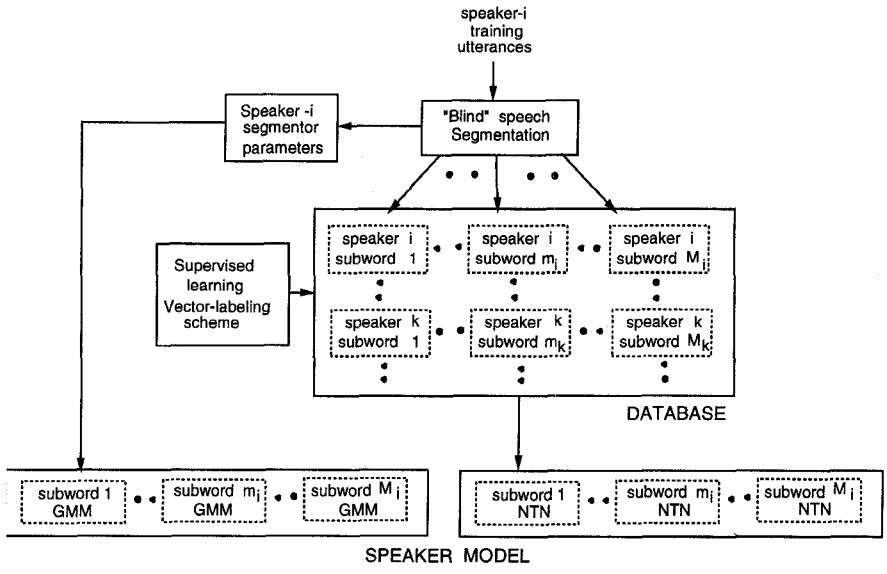

Figure 1: Training a speaker model

The procedure to train a model for a given speaker is illustrated in Figure 1. The multiple repetitions of the speaker's password are used by the segmentation module to estimate the number of subwords in the password along with the subword boundaries. The mean vector and diagonal covariance matrix of the subword segments are obtained as by-products of the segmentation module. These are used as the GMM component of the speaker model. For each subword segment of the password, a NTN model is also trained. The *closest* subword segments from other speakers who are already enrolled in the database are used as non-target data for training these subword NTN models.

The procedure to verify a claimed identity is illustrated in Figure 2. The given testing utterance is segmented to the optimal number of segments determined during training. The subword segment vectors are scored using the appropriate subword NTN and GMM models. The scores of these subword segments are averaged and a composite score for the entire phrase is obtained. The phrase-level NTN and GMM scores are then fused together using the linear opinion pool. We have performed experiments that did not show any advantages by combining this information at the subword level. If multiple models are obtained during training using the leave-one-out method, then all these models are scored in the above manner. These model scores are averaged to yield the final output score.

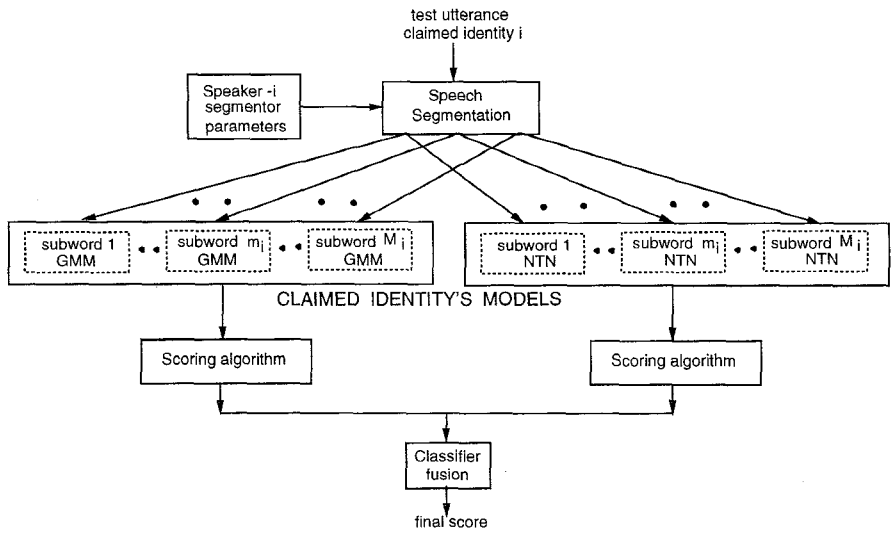

Figure 2: Testing a claimed identity

4 EXPERIMENTAL RESULTS

The T-NETIX *SpeakEZ Voice Print*^{SM} system is evaluated with three toll quality speech corpora that were collected by T-NETIX. The first database is known as the "names" database. The names database consists of 10 male target speakers, each with three enrollment utterances of their full name. The imposter attempts are comprised of 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 separate non-target 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 evaluation speakers and and 15 non-target speakers. The aspects of each database are summarized in Table 1. The non-target speakers column in Table 1 refer to the *development* set that is used during training of a speaker model. 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.

The first experiment evaluates the system equal error rate as a function of the number of segments. Generally, the system computes the number of segments per password, but in this case, we have forced the number of segments to be constant for all speakers. The results of this experiment as performed on the names database are shown in Figure 3. It is clear from Figure 3 that the GMM requires several segments before the performance

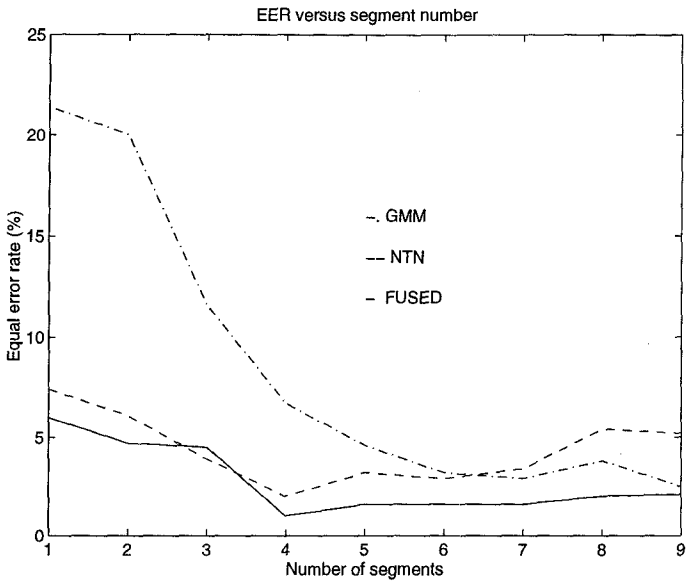

Figure 3: EER versus number of segments

starts to become competitive. The performance of the NTN, however, starts o degrade as the number of segments increases beyond four or five segments. This is due to the fact that the number of data samples per NTN decreases as he number of segments increases. Hence, for the NTN the lack of data starts o overcome the benefits of decomposing the acoustic space of the password.

The next experiment evaluates the equal error rate as a function of alpha or the linear opinion pool method of data fusion. The system uses a variable umber of segments per speaker. The results of this experiment for the names latabase are shown in Figure 4. Here, it can be seen that the individual erformance of the GMM and NTN is 3.2% and 3.4%, respectively. However, y combining the results of these methods, the EER can be reduced to 1.6%.

This experiment was also evaluated with the "Open Sesame" and "cellu-ar" database and the results for these experiments are shown in Figures 5 nd 6, respectively. The results for the "Open Sesame" database show the ndividual performance of the NTN and GMM to be 1.6% and 2.3%, respec-vely, whereas the performance of the fused output is 0.9%. For the cellular Al Capone" database the individual performance of the NTN and GMM is 1.8% and 10.2%, respectively, while the performance of the fused output is .2%.

The experimental results for T-NETIX's *SpeakEZ Voice Print*^{SM} system e tabulated for the "names", "Open Sesame" and "cellular" databases in able 1. The results in this table reflect the fusion results when $\alpha = 0.5$.

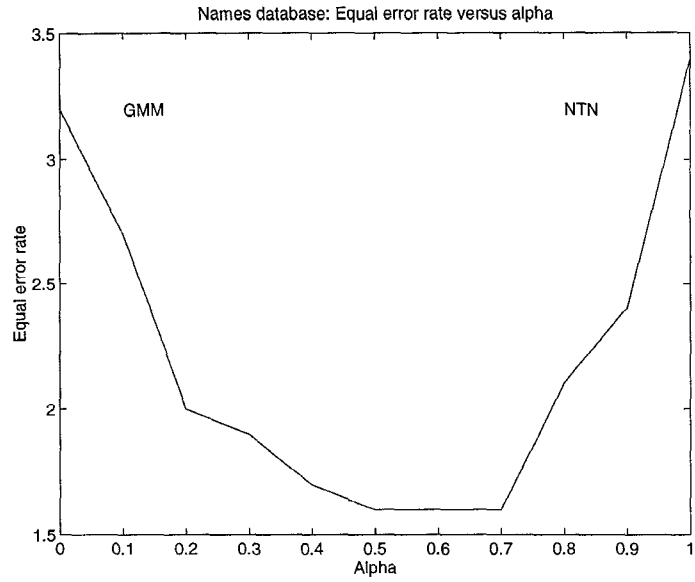

Figure 4: Linear opinion pool for names database

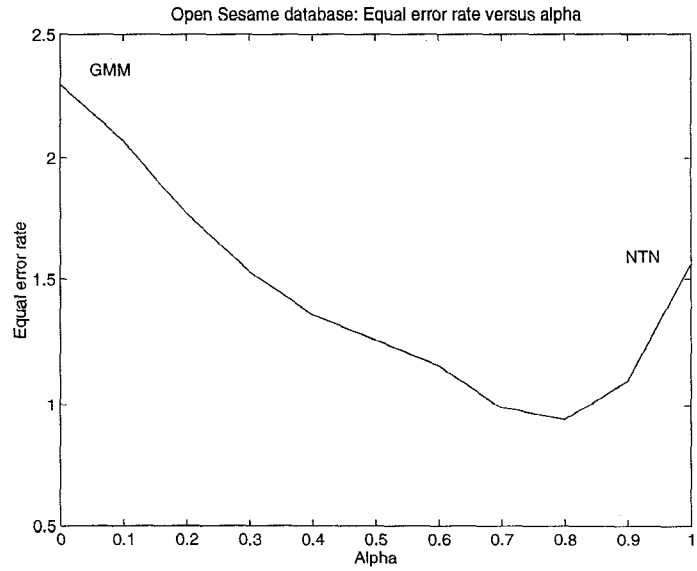

Figure 5: Linear opinion pool for "Open Sesame" database

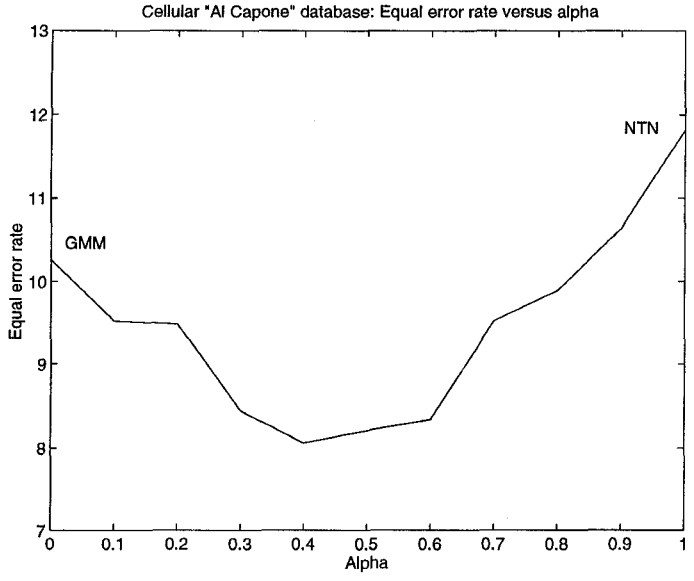

Figure 6: Linear opinion pool for "Cellular" database

5 CONCLUSION

The T-NETIX *SpeakEZ Voice Print*^{SM} system is evaluated for several text-dependent speaker verification tasks. These include applications in both cellular and landline environments. The T-NETIX *SpeakEZ Voice Print*^{SM} system does not have any constraints on the vocabulary from which the password is selected. This is accomplished through the use of sub-word neural tree networks and a blind segmentation algorithm that does not require phonetic label information. In addition, the system utilizes concepts within data fusion to capitalize upon different modeling approaches whose errors are uncorrelated. The data fusion techniques are found to reduce the error rate by a factor of two for the landline databases. The error rate for the cellular database is reduced by 20%. The error rates for the landline and cellular databases

Password text	# development/ evaluation speakers	# true/imposter trials	Performance (EER)
"Open Sesame"	47/56	195/11,229	1.3 %
Own full name	80/10 males	100/450	1.6 %
"Al Capone"	15/26	273/6825	8.2 %

Table 1: Performance for the *SpeakEZ Voice Print*^{SM} system

are roughly 1-2% and 8%, respectively. We find these results very encouraging given the constraints of limited training repetitions, short enrollment utterances, and unconstrained vocabulary for password selection.

References

[1] D. Naik. Pole-filtered cepstral mean subtraction. In *Proceedings ICASSP*, pages 157–160, 1995.

[2] A. Higgins, L. Bahler, and J. Porter. Speaker verification using randomized phrase prompting. *Digital Signal Processing*, 1:89–106, 1991.

[3] C.S. Liu, C.H. Lee, B.H. Juang, and A.E. Rosenberg. Speaker recognition based on minimum error discriminative training. In *Proceedings ICASSP*, pages 325–328, 1994.

[4] 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.

[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. Text-dependent speaker verification using data fusion. In *Proceedings ICASSP*, 1995.

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

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

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

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

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

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

[13] M. Sharma and R.J. Mammone. Blind speech segmentation: Automatic segmentation of speech without linguistic knowledge. In *Proceedings ICSLP*, 1996.