MODEL ADAPTATION METHODS FOR SPEAKER VERIFICATION

William Mistretta and Kevin Farrell
T-NETIX/SpeakEZ Inc.
67 Inverness Drive East
Englewood, Colorado 80112
email: bill.mistretta@t-netix.com

ABSTRACT

Model adaptation methods for a text-dependent speaker verification system are evaluated in this paper. The speaker verification system uses a discriminant model and a statistical model to represent each enrolled speaker. These modeling approaches consist of a neural tree network and Gaussian mixture model. Adaptation methods are evaluated for both modeling approaches. We show that the overall system performance with adaptation is comparable to that obtained by training the model with the additional information. However, the adaptation can be performed within a fraction of the time required to retrain a model. Additionally, we have evaluated the adapted and non-adapted models with data recorded six months after the initial enrollment. The adaptation reduced the error rate for the aged data by 40%.

1. INTRODUCTION

Speaker verification consists of determining whether or not a voice sample provides sufficient match to a claimed identity. Speaker verification has matured to the point where commercial deployments of the technology are now available. One critical aspect of a speaker verification system that can directly attribute to its success is robustness to intersession variability and *aging*. Intersession variability refers to the situation where a person's voice can experience subtle changes when using a verification system from one day to the next. A user can anticipate the best performance of a speaker verification system when performing a verification immediately after enrollment. However, over time the user may experience difficulty when using the system. For substantial periods of time, such as several months to years, the effects of aging may also degrade system performance. Whereas the spectral variation of a speaker may be small when measured over a several week period, as time passes this variance will grow [1]. For some users, the effects of aging may render the original voice model unusable.

Due to the effects of intersession variability and aging, models that are trained with data from a single enrollment session have a limited chance of success. Studies have been performed that evaluate feature robustness with respect to intersession variability and aging [1]. However, it was found that the models still needed to be trained with the data from several sessions to be effective. One approach to accommodate intersession variability is to have several initial enrollment sessions for each user in the system. This option is perhaps the most convenient from a technology standpoint, however, a burden is now placed upon the user. Another option that is less inconvenient to the user is to adapt the model with verification utterances that have passed some acceptance criteria. This paper considers the latter option.

Model adaptation methods have been explored extensively in the field of speech recognition. Some of these methodologies have been extended to applications in speaker recognition. For example, methods have been proposed to adapt speaker-independent hidden Markov models with data from a target speaker to create a model for speaker verification [2]. Methods have also been evaluated for adapting dynamic time warping (DTW) approaches by averaging the new observation with the original template [3]. These two modeling approaches are based on statistical and distortion measures, respectively. Far less attention has been devoted to adapting *discriminant*-based models that are trained with supervised training algorithms.

We propose a new adaptation scheme for a text-dependent speaker verification system. The speaker verification system uses both a discriminant and a statistical model to represent a user. Given a test utterance that belongs to the target speaker, both models are adapted with that utterance. The resulting performance after adaptation is comparable to that obtained by training the model with the original enrollment utterances in addition to the adaptation utterances. The adaptation process, however, can conveniently be performed following a verification while consuming minimal computational resources. An additional benefit of adaptation is that the original training data does not need to be stored, which can be burdensome for systems deployed within large populations.

The remainder of this paper is organized as follows. The following section provides a description of the modeling approach used in our speaker verification system. This is followed by a description of the adaptation methods that are used for each model component. Experimental results are provided for the adaptation methods in addition to results for data collected six months after the initial enrollment. A summary and conclusion are then provided.

2. SPEAKER VERIFICATION MODELING

The modeling approach used in this paper is based on subword modeling and data fusion. Speech is first segmented into sub-words using a blind segmentation algorithm [4]. The data at each sub-word is then modeled with a neural tree network (NTN) and Gaussian mixture model (GMM). The NTN provides a discriminative-based speaker score and the GMM provides one that is based on a statistical measure. The outputs of these two modeling approaches are combined using data fusion. Since these two modeling approaches tend to have errors that are uncorrelated, performance improvements can be obtained by combining the model outputs. The architecture of the system is illustrated

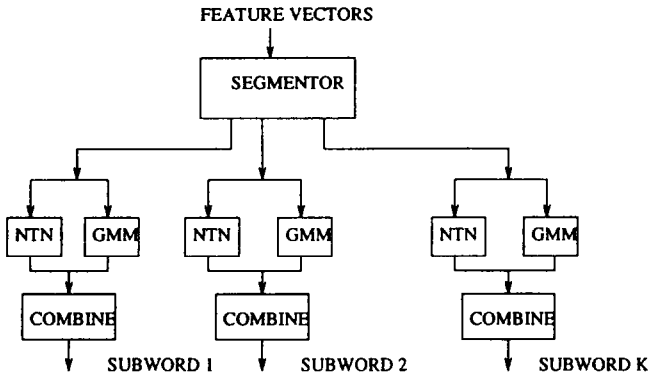

Figure 1. Speaker Verification Model

in Figure 1. The NTN, GMM, and data fusion method are now described in more detail.

2.1. Neural Tree Network

The NTN [5] 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 learns 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 correspond to leaves in the NTN that contain probabilities. These probabilities represent the likelihood of the target speaker having generated data that falls within that region of feature space [6]. The NTN has been evaluated for text-independent speaker verification [6], whole-word-based text-dependent speaker verification [7], and subword-based text-dependent speaker verification [8, 9].

2.2. Gaussian Mixture Model

The Gaussian mixture model (GMM) has been evaluated for numerous tasks within speaker recognition [10, 11]. Essentially, a region of feature space for a target speaker is represented by a set of multivariate Gaussian distributions. The GMM probability distribution function is expressed as

$$p(\mathbf{x}|\theta) = \sum_{i=1}^{C} P(\omega_i) p(\mathbf{x}|\mu_i, \sigma_i^2). \quad (1)$$

Each of the C mixture components is defined by a mixture weight $P(\omega_i)$ and normal distribution function $p(x|\mu_i, \sigma_i)$. The normal distribution is constrained to have a diagonal covariance matrix defined by the vector σ_i. The PDF is used to produce the sub-word GMM score. Scores are summed across sub-words to obtain a GMM model score for the phrase as a whole.

2.3. Data Fusion

In this paper, we use a linear opinion pool method to combine the output scores from the NTN and GMM. 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. Data fusion has been evaluated for combining models in several speaker verification applications. These include NTNs with DTW [7], NTNs with GMMs [9], and DTW with VQ [12].

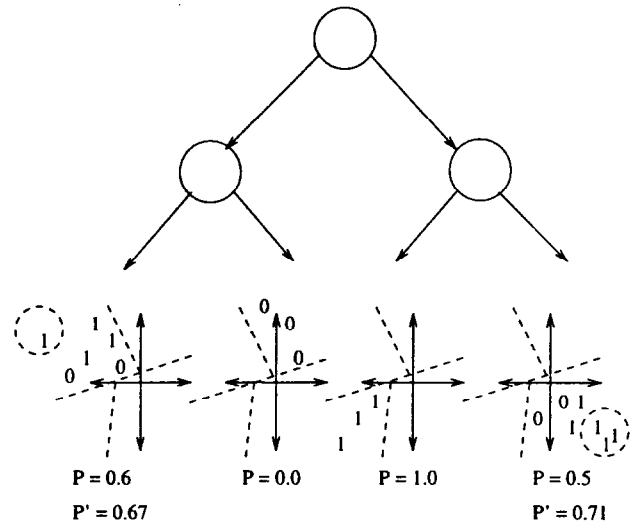

Figure 2. NTN Adaptation

3. ADAPTATION METHODS

Adaptation methods are provided for both model approaches described in the previous section. The adaptation occurs during verification. First, features are extracted for an adaptation utterance. These features are then segmented into sub-word partitions that can be processed by the corresponding NTN and GMM models at each subword. The adaptation methods for the NTN and GMM models are described in the following sub-sections.

3.1. NTN Adaptation

A NTN determines the speaker score for a given vector by traversing the tree and retrieving the probability at the leaf which the vector arrived. The probability at each leaf of the NTN is computed as the ratio of speaker observations to total observations encountered during training. By maintaining the number of speaker observations and impostor observations at each leaf, the probability update is straight-forward. Each vector of the adaptation utterance is applied to the NTN and the speaker observation count of the leaf that the vector arrives at is incremented. During testing, the probabilities are computed with the new leaf counts. This concept is illustrated in Figure 2 where the adaptation vectors are those within the dashed circles. For the left-most leaf in Figure 2, the original probability is computed as 0.6 and the adapted probability is 0.67. For adaptation utterances we have also found advantages by applying more weight to the new observations.

Since only the leaves of the NTN are modified during adaptation there is the implicit assumption that the feature space partitions do not have to change. Adapting the discriminant boundaries is not feasible as the nodes and leaves only retain information regarding the weight vectors and observation counts, respectively. If the discriminant boundaries must change, then retraining is the most practical solution.

3.2. GMM Adaptation

Each GMM is adapted individually using sub-word data acquired from the blind segmentation. The remainder of this section describes the adaptation of a single sub-word GMM since the process is identical for each sub-word. Referring to Equation 1, the adaptation process produces an updated set of GMM Parameters $\{P(\omega_i)', \mu_i', \sigma_i'^2; i = 1...C\}$ for the GMM PDF that reflects the contribution of the adaptation phrase.

A clustering of the adaptation data is performed as the first step in the individual GMM adaptation. If the adaptation features are defined by X with N vectors, the clustering groups the data into C subsets $\mathbf{X}^i; i = 1...C$, where $\mathbf{X}^i$ contains N_i vectors. A simple Euclidean distance between the input vector and component distribution means is used to partition the data.

The verification model retains information on the number utterances used to train the GMM along with the number of prior adaptations. The sum of these values M is used to scale the mixture weights, means, and variances before adding new statistics. The algorithm also makes the assumption that the prior utterances all contain N training vectors. It does this because the true sizes of the previous training and adaptation utterances are not retained as part of the verification model. Given these assumptions, the adapted component distribution parameters can be defined as follows:

$$P(\omega_i)' = \frac{P(\omega_i)MN + N_i}{(M+1)N}, \quad (3)$$

$$\mu_i' = \frac{\mu_i MNP(\omega_i) + \sum_{j=1}^{N_i} \mathbf{x}_j^i}{MNP(\omega_i) + N_i}, \quad (4)$$

and

$$\sigma_i'^2 = \frac{\sigma_i^2 M(N-1)P(\omega_i) + \sum_{j=1}^{N_i} (\mathbf{x}_j^i - \mu_i')^2}{M(N-1)P(\omega_i) + N_i - 1}. \quad (5)$$

This approach to adapting the distribution parameters weights all training and adaptation utterances equally. This means that each new adaptation phrase has less effect on the GMM. By limiting M to a maximum value, a simple forgetting factor can be incorporated into the adaptation. The forgetting factor was not examined for this paper.

4. EXPERIMENTAL RESULTS

All results discussed in this paper are produced from experiments conducted on a verification database that contains nine enrolled speakers. Additionally, there are 80 separate speakers that are used as the development speakers for training the neural tree network. The database contains two data sets with collections separated by a six month period. The first set contains 13 repetitions of each person speaking their full name and five repetitions of them speaking each other person's name. This amounts to 58 recordings for each speaker. The second set contains ten more repetitions of each person speaking their own name. We refer to a repetition of a person saying their own name as a true-speaker repetition and a person saying another person's name an impostor repetition. The two data collections are referred to as the *recent* set and *aged* set respectively.

Three training scenarios are examined for the paper. In each case, all training repetitions were taken from the recent collection set. The scenarios are outlined below.

1. Train a verification model with three true-speaker repetitions. (TR3)
2. Train a verification model with six true-speaker repetitions. (TR6)
3. Train a verification model with three true-speaker repetitions and adapt on three true-speaker repetitions. (TR3AD3)

For the second and third training scenarios, the first three training repetitions are kept fixed, while the second three repetitions are varied using a resampling scheme. The resampling technique is based on a leave-M-out data partitioning where M=3. For each training, three new repetitions are used. This allows for three independent trainings for the ten available true-speaker repetitions. The fixed training repetitions used for scenarios 2 and 3 are the same as those used in scenario 1. The first scenario provides a baseline system performance, the second shows the benefit of adding speaker information to the original training, while the third shows the benefit of adapting the model using the additional speaker information.

A set of three experiments are initially performed for each training scenario. This include testing the GMM and NTN models individually along with testing a combined model. All testing repetitions are taken from the recent collection set. For the baseline training scenario, ten true-speaker repetitions and 45 impostor repetitions are tested for each speaker model. Equal error rates (EER) are then calculated for the system by collecting performance across speakers. For scenarios 2 and 3, three resampling tests are performed for each individual experiment. For each test, the appropriate three true-speaker repetitions are excluded from the experiment. This results in 7 true-speaker and 45 impostor repetitions for each test or 21 true-speaker and 135 impostor repetitions for each speaker.

Table 1 displays the performance of these experiments. Several observations can be made when inspecting the table. First, the additional speech data provides a performance benefit when the model is trained on all the data. Second adapting on the additional training data also improves performance to some degree. The GMM adaptation does a better job at matching the training performance than the NTN adaptation. Although the NTN does not adapt as well as the GMM, it still helps reduce the EER when applying adaptation to the combined model.

	Verification Model Type		
Training Senerio	GMM	NTN	Combined
TR3	5.3%	6.0%	4.0%
TR6	1.9%	1.8%	0.63%
TR3AD3	1.7%	4.3%	1.5%

Table 1. Verification EER performance for several training senerios and verification model types. All experiments evalulated with the *recent* collection data.

A second set of experiments are performed for the combined verification model. For this set, true-speaker testing repetitions are taken from the aged collection set. All other training and testing conditions are kept the same as the previous experiments. These results are displayed in Table 2. The table shows that all training scenarios suffer when evaluating the aged true-speaker repetitions. This is to be expected, since the verification model is trained on data collected over a short period of time. There is still improvement though when the model is trained on additional data

from the recent set. As with the previous experiments, the adaptation also improves the performance but not as much as the full training.

Training Senerio	Combined Model
TR3	12.%
TR6	5.4%
TR3AD3	7.2%

Table 2. Verification EER performance for several training senerios and combined model type. All experiments evalulated with the *aged* collection data.

5. CONCLUSION

This paper examines model adaptation methods for a text-dependent speaker verification system. Adaptation techniques are examined for both a GMM and NTN. It was shown that GMM performance improved from 5.3% to 1.7% and NTN performance improved from 6.0% to 4.3% when adapting on additional training data. A classifier that combines these two models shows similar improvement and performs better than either classifier in isolation. In addition, when testing the combined classifier on aged data, the performance improves from 12.% to 7.2%. The overall system performance using adaptation is comparable to that achieved by training the model with the adaptation information.

REFERENCES

[1] S. Furui. Comparison of speaker recognition methods using statistical features and dynamic features. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-29:342–350, April 1981.

[2] T. Matsui and S. Furui. Speaker adaptation of tied-mixture-based phoneme models for text-prompted speaker recognition. In *Proceedings ICASSP*, pages 1125–1128, 1994.

[3] J.M. Naik and G.R. Doddington. High performance speaker verification using principal spectral components. In *Proceedings ICASSP*, pages 881–884, 1986.

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

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

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

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

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

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

[12] J. Schalkwyk, N. Jain, and E. Barnard. Speaker verification with low storage requirements. In *Proceedings ICASSP*, 1996.