INTEGRATION OF SPEAKER AND SPEECH RECOGNITION SYSTEMS

D. A. Reynolds and L. P. Heck
Dept of Electrical Engineering
Georgia Institute of Technology
Atlanta, GA 30332

ABSTRACT

This paper presents a novel combination of a high-performance speaker identification system and an isolated word recognizer. The front-end text-independent speaker identification system determines the most likely speaker for an input word. The speaker identity is then used to choose the reference word models for the speech recognizer. When used with a closed set of speaker, the combination is a system capable of automatically producing speech and speaker identification. For an open set of speakers, the speaker recognition system acts as a "speaker quantizer" which associates the unknown speaker with an acoustically similar speaker. The matching speaker's word models are used in the speech recognizer. The application of this front-end speaker recognizer is described for a DTW and HMM speech recognizer. Results on a combination using a DTW word recognizer are 100% for closed set experiments. Open set results are 92.6%; an increase of 11.4% from cross speaker word recognition rates and comparable to speaker-dependent performance.

1 INTRODUCTION

Despite the similarity between the areas of speech and speaker recognition, relatively little research has been done in combining these areas. The speech signal is a complex information bearing signal which can be considered to be comprised of two major components: the underlying text and the characteristics of the speaker. Traditionally, speaker recognition is concerned with determining the source of the speech signal, considering the phonetic variations to be extraneous noise, whereas speech recognizers are aimed at minimizing the speaker dependent variations in the speech signal in order to concentrate on the textual component. However, due to the complicated integration of these components, a better approach is to combine these systems in such a way that as much information as possible is extracted and used for the final task. One problem which is well suited for this integration is improving speaker independent performance in speech recognition systems. This combination would have applications not only in improving speaker independence in speech recognizers but also in producing a system which simultaneously recognizes spoken text and speaker identity.

As speech recognition systems aim toward large vocabularies, it is infeasible to train the system for each new speaker, so speaker independent system are required. In order to achieve speaker independence, present speech recognition systems employ several methods which are somewhat similar to a crude speaker recognition system. Clustering techniques which produce multiple models per word attempt to find representative exemplars for groups of similar sounding speakers. However, this training uses all the training speakers' data in the unsupervised clustering so that the speaker differences are "blurred" in the final word models and generally perform worse than using unclustered models derived from an individual speaker [1]. Using all reference speakers' word models preserves individual differences but produces a computationally expensive search during recognition.

To alleviate these problems, this paper presents the use of a speaker recognition system as a front end processor to a speech recognition system . This work is based on a statistical speaker recognizer capable of recognition rates >90% for 1 second test utterances [2]. The combination operates under two situations. With a closed set of speakers the system produces simultaneous speaker and speech recognition. For an open set of speakers, the front-end system acts as a "speaker quantizer" which matches a new speaker to a reference word model from an acoustically similar training speaker. The speech recognizer then uses the associated reference speaker's word templates to perform word recognition. In this way the word templates are automatically adapted to any new speaker. The combination is modular and can be used with any speech recognition system. The use of the speaker recognizer with a DTW isolated word recognition system is examined and its with a HMM speech recognizer is described.

The rest of the paper is organized as follows: The next section presents the Gaussian Mixture Model (GMM) speaker recognition system. A description of the use of the speaker recognizer with a DTW and HMM system follows. Then results from some experiments using a combined GMM-DTW system are presented.

2 SPEAKER RECOGNITION SYSTEM

A robust text-independent speaker identification system capable of high recognition rates for utterance lengths as short as 1 second was introduced in [2]. This system, based upon the use of Gaussian mixture densities to statistically characterize speakers, acts as a hybrid between two effective models for speaker identification: uni-modal Gaussian classifiers [3] and vector quantizer codebooks [4]. The Gaussian mixture model (GMM) combines the robustness of the parametric Gaus-

CH2977-7/91/0000-0869 $1.00 © 1991 IEEE

sian model with the short utterance performance of the non-parametric VQ model.

Similar to a uni-modal Gaussian classifier, the GMM classifier represents each speaker by a pdf governing the distribution of his/her feature vectors. The Gaussian mixture model has the form

$$p(\vec{x}|\Theta) = \sum_{i=1}^{C} P(\omega_i)\, p(\vec{x}|\omega_i). \quad (1)$$

The density is a weighted linear combination of C component uni-modal Gaussian densities, $p(\vec{x}|\omega_i)$, each parameterized by a mean vector and covariance matrix, $\vec{\mu}_i$ and Σ_i (collectively denoted as Θ). The ith component has the interpretation of representing a "hidden" acoustic class ω_i as shown in Figure 1.

The structure of this model is quite flexible in form and has several strong points. The first is the mixture model's ability to form smooth densities of irregular shape. Secondly, the GMM is a general parametric model that can take on various forms. As shown, the GMM has a full covariance matrix per mixture component, but the model can have diagonal covariance matrices, a global covariance for all components, or even a fixed covariance matrix for all speaker models. The decision of what form the model should take is usually an empirical trade-off between the number of free model parameters and the amount of training data available. In this work speaker models use 10 Gaussians per mixture each with a diagonal covariance matrix. Also, due to its form, the GMM can be made quite robust to noise and channel degradations [2, 5, 6].

In Eq (1) only the number of classes C is assumed to be known while all other parameters must be estimated simultaneously. Given an appropriate starting point, the iterative Expectation Maximization (EM) algorithm [7] is used to find the maximum likelihood estimates of the density parameters. Training generally requires only 5–10 iterations. The models are trained using the same data used for the speech recognizer, so only one training data collection is required.

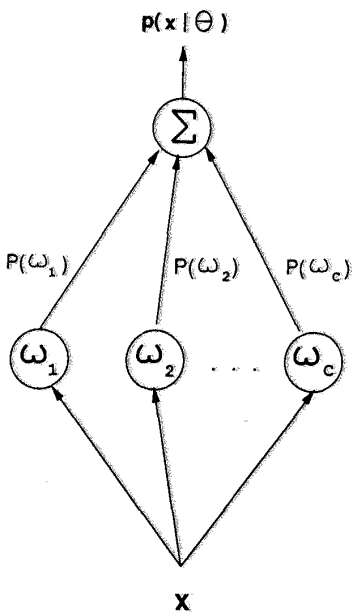

Figure 1: Gaussian mixture speaker model. Acoustic classes, ω_i, are represented by Gaussian distribution.

It is also possible to introduce discriminant training into the GMM training procedure. Referring to Figure 1, note that the GMM has a similar topology to a radial basis function network, which allows for using similar discriminant training methods. First we obtain maximum-likelihood estimates of the GMM parameters using the EM algorithm. The mixture weights are then replaced by discriminant weights derived from all the speaker's data. The discriminant weight training is non-iterative requiring only a single matrix inversion [8]. While discriminant training is quite important for large speaker populations with many similar sounding speakers, for the small speaker set employed in this work the single stage maximum-likelihood training method was sufficient.

For classification, the likelihood function for each speaker's model is computed over the input word and the speaker model producing the highest likelihood score indicates the chosen speaker.

3 SPEAKER QUANTIZATION

3.1 Application to DTW

In Figure 2 a tandem connection of the GMM speaker recognition system and the DTW word recognizer is shown. The system operates as follows: After the input utterance is processed to produce a stream of feature vectors (20th order mel-cepstral vectors in this work), the most likely speaker $\hat{s}$ is determined by the GMM speaker identification subsystem. The word templates for the estimated speaker are selected from the super-set of all speaker's templates and passed on to the DTW word recognition system. The word recognizer then determines the unknown word in a normal fashion using the estimated speaker's reference templates. If the whole system is run with a closed set of speakers, then both the estimated word and speaker identity are final outputs.

Since the DTW word recognizer does not require any model training, the system training is straightforward. First, as with other speaker independent methods, a representative set of speakers should be selected. This set should cover all expected speakers in the final user population. However, if the system is to be used for speaker and word identification, then the training speakers need only be those of the closed reference set. Next, reference word templates for each speaker are collected to produce the DTW super-set reference codebook. These same word references are then used to train each speaker model via the EM algorithm. For a vocabulary size of more than 10 words, there should be ample data per speaker for training the speaker models. Since the system uses linguistically constrained speech, the speaker models do not need a large amount of data to train the acoustic classes.

This system can be viewed as performing "speaker quantization" prior to speech recognition. The GMM speaker recognizer will associate a new speaker to a closely matching reference speaker in much the same way a vector quantizer associates a vector to the best match in its codebook. Thus a large reference model codebook can be maintained and the speaker recognizer acts to direct the codebook search during recogni-

tion. This approach is similar to work on speaker clusters [9] and speaker hierarchical clustering [1].

One of the main advantages to this approach is that it allows a relatively simple way to maintain a large set of reference models that cover many different speaker acoustical variations while not incurring the large computational cost of searching through all models during word recognition. After speaker identification, the recognition cost is the same as a single model system.

3.2 Application to HMM

The speaker-speech recognizer combination can also be used with a HMM word recognizer in the same manner as with the DTW word recognizer. For a whole word left-to-right HMM, each word model now consists of the transition probabilities and the state densities rather than a collection of reference templates. The speaker recognizer then performs as above, by determining the most likely speaker and directing the word recognizer to use that speaker's word models.

However, since the both the HMM and the GMM are statistically based classifiers, a more interesting combination is for the speaker and speech recognizer to use the same probability models. Both the GMM and HMM are trained on the same data to optimize the ML criteria, so their underlying densities should be similar. Whereas the GMM is less constrained than the HMM, the HMM is a generalized model and it should be possible to use the HMM state densities as the acoustic class densities in the GMM. The idea is to train each speaker's HMM word models as normal and then form a speaker's GMM by pooling the state densities over all of his/her word models. The weights of the GMM could then be trained using the discriminant method described in Section . Clustering of similar component densities would be applied if there were too many components in the GMM (Models with S states and W words in the vocabulary would give $S \times W$ component Gaussians.) This pooling approach takes advantage of the similarities between the speaker and speech models and allows a form of model sharing which can reduce storage and training time.

The speaker recognizer is also easily adapted for use in a continuous speech recognition. Using a HMM continuous speech recognizer, the speaker recognizer would select the appropriate sub-word models to be used out of a collection of speaker-dependent models. The above pooled training approach could also be used with the sub-word models. The speaker recognizer would be used in a continuous mode wherein it produces a time-varying likelihood score for each speaker model. The likelihood scores are then used to label segments of speech as to speaker identity.

4 RESULTS AND DISCUSSION

In this section, the results from experiments using a combined GMM speaker recognizer and DTW isolated word recognizer are presented. These preliminary experiments were performed to evaluate the effectiveness of using the combined system and were carried out using the following setup:

Sample rate: 8000 Hz

Features used: 20th order FFT derived mel-frequency cepstral vectors

Number of speakers: 5 (3 male, 2 female)

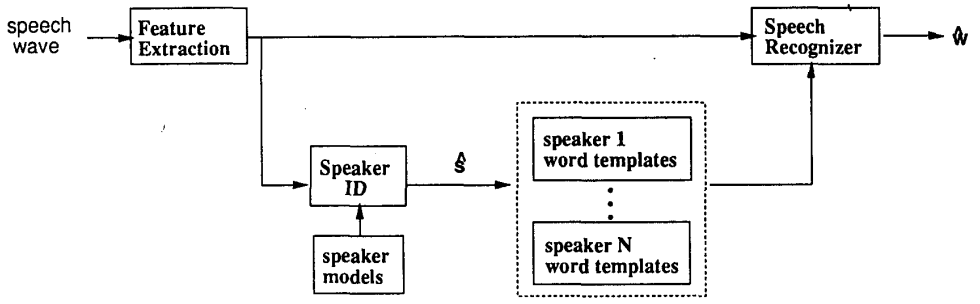

Figure 2: Combined speaker-speech recognition system. GMM speaker identification system determines estimated speaker, $\hat{s}$, which is used to select the word models used by the speech recognizer.

Vocabulary: isolated digits (0–9)

Training data: 5 tokens per word per speaker

Testing data: 5 tokens per word per speaker

The speakers are identified by initials: dar, kpn, rhb (male) and kcg, uhh (female). The speaker GMMs consisted of 10 Gaussian components with a diagonal covariance matrix per component and were trained using five digit sets (approx 20 sec).

The first experiment examined how the combined system performed for a closed set of speakers (i.e., digit templates and speaker models were available for all input speakers). Using the speaker recognizer front-end, 100% digit and speaker recognition was achieved, which is in agreement with 100% digit recognition using speaker-dependent reference templates. With perfect speaker recognition, this result is not surprising. The main result was that speaker-dependent performance was achieved without knowing the input speaker's identity and without searching through all speaker's reference templates. An experiment using a large set of speakers and examining performance with speaker misclassifications would be the next step to give a better picture of system performance under more rigorous conditions.

Next, to evaluate the ability of the speaker quantizer to match a new speaker with an acoustically similar speaker, the combined system was operated using an open set of speakers (i.e., the input speaker's digit templates were not available and the speaker recognizer had to choose the best matching speaker). The results are shown in Table 1. For comparison, results are shown from cross speaker recognition, where speaker A's input digits are recognized using speaker B's reference templates. The results in the table are the average of using each speaker as the reference speaker. The cross speaker results represent the average DTW performance when a single speaker is chosen as the reference speaker. It is clear that the speaker quantizer improves on this recognition rate; increasing the average digit recognition rate by 11.4%. It is also seen that the speaker quantizer compares favorably to speaker-dependent performance.

	Cross Speaker	Speaker Quantization (open set)
dar	81.5	97.0
kcg	74.5	79.0
kpn	83.5	100.0
rhb	88.5	96.0
uhh	78.0	91.0
AVG	81.2	92.6

Table 1. Digit recognition results (in %) for all speakers.

This increased performance is believed to be due to the system's ability to choose different reference models for word recognition for each new input utterance. The reference templates are not set, but are adaptively changed according to the input utterance, thus allowing for cases when an input speaker's pronunciation on different words matches different speakers. In fact, for the above experiment, reference templates for a test digit were often from different speakers for different occurrences of the test digit. A notable exception was that the two female speakers always matched to each other. However, the open set results highly depend on a complete speaker population to ensure a new speaker will have an acoustically close match.

5 CONCLUSION

A novel combination of a speaker recognizer and a word recognizer was presented in this paper. This combination can be used as an automatic speaker and speech recognition system or as a means to associate an input speaker with an acoustically similar reference speaker to improve the speech recognizer's speaker-independent performance.

Using a DTW word recognizer in the combination, both closed and open speaker sets were examined for the isolated digits. Perfect speaker and digit recognition was achieved for the closed set task. Although not surprising for this small speaker set, the result demonstrates that the speaker recognizer can be used to reduce the template search space and still produce speaker dependent performance. Future experiments will focus on working with a larger speaker set. For an open speaker set, the speaker quantizer improved digit recognition performance by 11.4% when compared to cross speaker digit recognition. The speaker quantizer recognition rate of 92.6% also compares favorably to the speaker dependent recognition rate. These result are promising and point to the benefits of a tighter link between the areas of speech and speaker recognition.

REFERENCES

[1] L. Mathan and L. Miclet, "Speaker hierarchical clustering for improving speaker-independent hmm word recognition," ICASSP, pp. 149–152, IEEE, 1990.

[2] R. C. Rose and D. A. Reynolds, "Text-independent speaker identification using automatic acoustic segmentation," ICASSP, IEEE, 1990.

[3] R. Schwartz, S. Roucos, and M. Berout, "The application of probability density estimation to text–independent speaker identification," ICASSP, pp. 1649–1652, IEEE, 1982.

[4] F. Soong *et al.*, "A vector quantization approach to speaker recognition," ICASSP, pp. 387–390, IEEE, 1985.

[5] D. A. Reynolds, R. C. Rose, and M. J. T. Smith, "A mixture modeling approach to text-independent speaker identification," JASA (Suppl 1), vol. 87, p. s109, 1990.

[6] R. C. Rose, J. A. Fitzmaurice, and D. A. Reynolds, "Robust speaker identification in noisy environments using noise adaptive speaker models," ICASSP, IEEE, 1991.

[7] G. McLachlan, *Mixture Models.* New York, NY: Marcel Dekker, 1 ed., 1988.

[8] S. Renals and R. Rohwer, "Phoneme classification experiments using radial basis functions," IJCNN, pp. 461–468, 1989.

[9] K. F. Lee, *Large-vocabulary speaker-independent continuous speech recognition: the SPHINX system.* Ph.D. thesis, Carnegie Mellon University, 1988.