S51.10

TEXT INDEPENDENT SPEAKER IDENTIFICATION USING AUTOMATIC ACOUSTIC SEGMENTATION [1]

Richard C. Rose and Douglas A. Reynolds

Lincoln Laboratory, MIT
Lexington, MA 02173–9108

ABSTRACT

This paper describes an acoustic class dependent technique for text independent speaker identification on very short utterances. The technique is based on maximum likelihood estimation of a Gaussian mixture model representation of speaker identity. Gaussian mixtures are noted for their robustness as a parametric model and their ability to form smooth estimates of rather arbitrary underlying densities. Speaker model parameters are estimated using a special case of the iterative Expectation–Maximization (EM) algorithm [4], and a number of techniques are investigated for improving model robustness. The system was evaluated using a 12 reference speaker population from a conversational speech database, and achieved 89% average text independent speaker identification performance for a *1 second* test utterance length.

1 INTRODUCTION

In the text independent speaker identification paradigm addressed in this paper, the speaker identification system is presented with a short utterance from an unknown speaker and must identify the unknown speaker as a member of a closed set of reference speakers. Reference speakers are represented by training data that is labeled only according to speaker identity, but completely unlabeled and unconstrained according to word or acoustic class. In general, input utterances to the speaker identification system are also completely unconstrained, and it is assumed that the identity of the speaker could change frequently in time. The interest here is in developing reliable speaker identification techniques that can identify an unknown speaker based on input utterances of length as short as 1 second. In these applications, formal enrollment procedures for the reference speakers are often not possible, so that speaker models must be trained on relatively short utterances (less than one minute).

There is already a large body of work applying statistical classifiers to the problem of text independent speaker identification. Existing approaches to text independent speaker identification rely on long–term spectrum averages or unsupervised clustering procedures to form statistical model representations of speakers. The features used in these systems are vectors representing the short–time spectral envelope of the speech signal. Two particularly effective models for speaker identity are the parametric Gaussian classifier, and the nonparametric vector quantizer. The multivariate Gaussian density has been shown to be an effective parametric classifier for speaker identification in a telephone speech paradigm [2]. The kmeans clustering algorithm was used to train nonparametric densities in the form of speaker dependent vector codebooks for speaker identification on a connected digit task [6]. Much of this work has been reviewed in [5].

As in all classification problems, the choice of the classifier is heavily dependent on the application. With proper precautions, continuous unimodal Gaussian distributions are known to exhibit good noise robustness properties; however, these models are formed from long–term spectrum averages and need reasonably long utterances for training and classification. Non–parametric speaker models have the benefit that they do not rely on an assumed parametric form for the underlying parameter densities. However, they are generally less robust, and classification performance tends to degrade rapidly under varying channel conditions.

This paper describes the use of the multimodal Gaussian mixture density for speaker identification. The parameters of this density are estimated using an unsupervised iterative maximum likelihood technique which is a special case of the Expectation Maximization (EM) algorithm [4]. There are several issues that motivate the use of Gaussian mixtures. First, Gaussian mixtures have properties of both the unimodal Gaussian distribution and non–parametric distributions. They share the robustness associated with the parametric Gaussian density, while at the same time share the ability of non–parametric models to model non–Gaussian distributed data. Second, as is illustrated in Section 6, Gaussian mixtures have a remarkable ability to form smooth parametric densities of irregular shape. Finally, the EM algorithm is intuitively appealing because it treats the underlying mixture components as representing "hidden" acoustic classes. It has long been postulated that some sort of acoustic class dependent representation of speaker identity is more appropriate than a simple long-term spectrum average, but it has not been clear what these acoustic classes should represent.

Following this introduction, the paper proceeds by describing the modeling techniques and the experimental paradigm, then presents experimental results. The procedure for maximum likelihood estimation of speaker dependent Gaussian mixtures is outlined in Section 2. Spectral normalization techniques for explicit compensation of speaker models with respect to speaker and channel variabilities are described in Section 3. Section 4 describes the experimental paradigm for evaluating the algorithm, and Section 5 describes the experimental results. Finally, Section 6 provides some discussion on mixture densities, with an overall summary provided in Section 7.

2 ESTIMATION of SPEAKER DEPENDENT MODEL PARAMETERS

In text independent speaker identification, it is assumed that the parameters of a statistical model for a speaker are trained from completely unlabeled observations taken from a representative utterance. The observation vectors, x_k, in this work are mel–frequency cepstra. It is assumed that a speaker ℓ is to be represented by a parameter vector, Θ^ℓ, where Θ^ℓ defines a Gaussian mixture density,

$$p(x_k|\Theta^\ell) = \sum_{i=1}^{C} P^\ell(\omega_i) p(x_k|\omega_i^\ell, \mu_i^\ell, \sigma_i^\ell) \,. \qquad (1)$$

The density is a weighted linear combination of C component unimodal Gaussian densities, $p(x_k|\omega_i^\ell, \mu_i^\ell, \sigma_i^\ell)$, where the ith component Gaussian represents a hidden state, ω_i, and is in turn represented by a mean vector and diagonal covariance matrix, μ_i^ℓ and σ_i^ℓ. While Equation 1 assumes a diagonal covariance matrix for each mixture component, in fact the form of the component Gaussians is an empirically derived tradeoff between model specificity and finite training data. These is-

[1] This work was sponsored by the Department of the Air Force.

sues are addressed in detail in Section 5. The component weights or mixing proportions, $P^\ell(\omega_i)$, and the component means and variances are all considered unknown, and must be estimated simultaneously. It is assumed that C, the number of mixture components, is known. Hence, the speaker dependent parameter vector is given as

$$\Theta^\ell = \{P^\ell(\omega_i), \mu_i^\ell, \sigma_i^\ell\}, \ i = 1, \ldots, C \tag{2}$$

Maximum likelihood estimates of Θ^ℓ are obtained by maximizing the log likelihood of a sequence of N observations,

$$L = \sum_{k=1}^{N} \log p(x_k|\Theta^\ell) \,. \tag{3}$$

Direct maximization of the above expression for L yields singular solutions, but given some initial estimate of Θ^ℓ, there exists a simple hill climbing procedure for obtaining a local maximum of L. The procedure, a special case of the EM algorithm, can be outlined as follows [4]:

Initialization: Begin with initial parameter estimates

$$P^\ell(\omega_i)^{(0)}, \mu_i^{\ell(0)}, \sigma_i^{\ell(0)}, \ i = 1, \ldots, C$$

1) Expectation: Compute the probability that the observations, x_k, $k = 0, \ldots, N-1$ belong to each hidden state, ω_i, $i = 1, \ldots, C$. Simply by applying Bayes rule to Equation 1,

$$p^\ell(\omega_i|x_k, \mu_i^{\ell(0)}, \sigma_i^{\ell(0)}) = \frac{p(x_k|\omega_i, \mu_i^{\ell(0)}, \sigma_i^{\ell(0)}) P^\ell(\omega_i)^{(0)}}{p(x_k|\Theta^{\ell(0)})} \,. \tag{4}$$

2) Maximization: Obtain maximum likelihood estimates of mixture parameters of Equation 2. Maximizing Equation 1 with respect to the model parameters yields an expression which depends on the initial parameter estimates through Equation 4. For example, the ith component mean vector is given by

$$\mu_i^\ell = \frac{\sum_{k=0}^{N-1} p^\ell(\omega_i|x_k, \mu_i^{\ell(0)}, \sigma_i^{\ell(0)}) x_k}{\sum_{k=0}^{N-1} p^\ell(\omega_i|x_k, \mu_i^{\ell(0)}, \sigma_i^{\ell(0)})} \tag{5}$$

3) Iterate: Repeat steps 1 and 2 with the initial parameter estimates replaced by those estimated in step 2).

Note that the component mean is simply a weighted linear combination of the training samples, where the weights are simply the a posteriori probabilities of each data vector occupying the associated class. The choice of an initial value for Θ^ℓ, $\Theta^{\ell(0)}$, can effect the final solution and the rate of convergence to a final solution. In Section 5 the performance of the speaker ID system is compared for different initialization strategies: one that performs an initial partitioning of the training data into acoustic classes, and another that performs a random partitioning of the training data.

It is also possible to modify the maximum likelihood Gaussian mixture model to discriminate among a closed set of reference speakers. This can be accomplished through reestimation of the mixture weights. The ML procedure estimates the $P^\ell(\omega_i)$'s in Equation 1 from the posterior probabilities of component membership given by Equation 4 as

$$P^\ell(\omega_i) = \sum_{k=0}^{N-1} p^\ell(\omega_i|x_k, \mu_i^\ell, \sigma_i^\ell) \,. \tag{6}$$

A discriminant speaker classifier could be designed so that the probability of an unlabeled observation for a particular speaker is dependent on a weighted linear combination of component Gaussians from all speakers. The component Gaussians can be trained from the above ML procedure, but the weights would be trained using a maximum a posteriori criterion based on representative observations from all speakers. A similar technique has been shown to provide improved speech recognition performance in an E–set phoneme recognition problem [3].

3 SPECTRAL NORMALIZATION

In using long utterances for training statistical speaker models, it is hoped that the reference speaker models will assimilate sources of speaker variability and phonetic variability so that what remains is a robust characterization of the speaker. It is well known, however, that there are significant sources of variability arising from differences in communications channels and intersession speaker variation.

The most direct approach for dealing with these sources of variability is to transform all reference speaker models based on unlabeled observations from the test utterance. It is assumed that all variability can be represented as a fixed linear component, and spectral normalization therefore be implemented using an additive bias vector in the cepstrum domain. Hence, reference speaker ℓ can be represented by the Gaussian mixture density, $p(x_k|\Theta^\ell, b^\ell)$, where b^ℓ represents the cepstrum bias vector. The bias vector affects the original mixture density in Equation 1 by a simply biasing the means of the component Gaussians,

$$p(x_k|\omega_i^\ell, \mu_i^\ell, \sigma_i^\ell, b^\ell) = \frac{1}{\sqrt{2\pi} \prod_{j=0}^{J-1} \sigma_{i,j}^\ell} \exp\left\{ -\frac{1}{2} \sum_{j=0}^{J-1} \left(\frac{x_{k,j} - \mu_{i,j}^\ell - b_j^\ell}{\sigma_{i,j}^\ell} \right) \right\} \tag{7}$$

Two techniques are investigated for estimating b^ℓ. The first is a form of blind deconvolution. In this case b^ℓ is computed as the difference between the average cepstrum of the input utterance and the average cepstrum of the training utterance. The second technique is essentially a model dependent compensation strategy, and is motivated by the speaker adaptation technique of Cox and Bridle [1]. If the underlying mixture classes are assumed to be "hidden" with respect to the cepstrum bias vector, the iterative algorithm described in Section 2 can be used to estimate the cepstrum bias vector. The component Gaussian probabilities in Equation 1 are replaced by Equation 7, and only b^ℓ is reestimated at each iteration as

$$b^\ell = \frac{1}{N} \sum_{k=0}^{N-1} \left(x_k - \sum_{i}^{C} p(x_k|\omega_i^\ell, \mu_i^\ell, \sigma_i^\ell, b^\ell) \mu_i^\ell \right) \tag{8}$$

It is not clear, however, whether allowing the probabilistic alignment of the observations with respect to the mixture classes to change in this manner will be beneficial to classifier performance. In Section 5, both procedures for estimating b^ℓ are evaluated in terms of their effect on average speaker identification performance.

4 DATABASE and EXPERIMENTAL DESIGN

The speaker identification experiments were performed on a 12 reference speaker population from a conversational database and consisted of 8 males and 4 females. Speaker models were trained using unlabeled utterances from each reference speaker reading a short paragraph. The speaker ID system was evaluated using excerpts from extemporaneous conversations involving all speakers. All data was recorded under studio conditions using an Electret condenser microphone mounted in a telephone handset and was sampled at 10kHz.

Throughout these experiments, twentieth order mel-frequency cepstrum vectors with the c[0] term removed were used as observations. The observation vectors were updated over 10 millisecond intervals and a front end energy thresholder was used to discard silence frames for both training and testing. With silence removed, training utterances ranged in length from 38 to 50 seconds per speaker. A preliminary experiment was performed to investigate the sensitivity of the training procedure to these differences in training utterance length. When the training utterance length was fixed at exactly 30 seconds for each speaker, no significant decrease in average speaker identification performance was observed over the observed performance for the full training set. So the differing amounts of training data do not in this case detract from overall performance.

For evaluation, the test conversations (after silence removal) were divided into fixed length blocks and the log likelihood of the observations in that block were computed for each speaker model. The block lengths used for likelihood score computation range in length from 0.01 to 5.0 seconds. The reported results represent the average correct speaker identification accuracy for approximately 2 minute utterances from all of the 12 speakers.

A baseline system was constructed using 50 component speaker dependent Gaussian mixtures with a common diagonal covariance matrix. The parameters were estimated via the EM algorithm using 20 dimensional cepstra vectors and a random initialization procedure that is discussed in the next section. In the following section several experiments using this baseline system and variants of the baseline system are discussed.

5 EXPERIMENTAL RESULTS

Using the experimental paradigm described above, a study was performed to evaluate the following issues in terms of their effects on the performance of the Gaussian mixture based speaker ID system. The first issue concerns the effectiveness of the multimodal Gaussian mixture parametric representation of a speaker density relative to the more widely used unimodal full covariance Gaussian model obtained from a long-term spectrum average. The Gaussian mixture requires considerably higher computational complexity in both model training and classification, so it is important to show that the multimodal representation can achieve higher classification accuracy.

The second issue addressed in the study was the iterative EM training algorithm: parameter initialization, number of iterations, and number of mixture components. The performance of the speaker ID system is compared for different initialization strategies: one that performs an initial partitioning of the training data into predefined acoustic–phonetic classes, and another that performs a random partitioning of the training data. The rate of convergence of the EM algorithm and how speaker classification performance is related to the number of mixture components are also examined. Finally, two techniques for compensating against some of the variabilities found between testing and training data are evaluated. These techniques include the cepstrum bias procedures described in Section 3, and the use of "invariant" features such as difference mel–frequency cepstra. It is interesting to note that, since the channel conditions are constant across all recordings in the database, the effect of the spectrum bias procedure is not to compensate for varying channel effects but instead to compensate for intersession speaker differences.

5.1 Baseline Performance

The performance of the baseline system and a unimodal Gaussian classifier are shown in figure 1. The unimodal Gaussian system is characterized by a 20 dimensional mean vector and a full covariance matrix. The figure shows that the mixture density based system outperforms the unimodal system for all test utterance lengths. Analysis of the percent identification accuracy for each speaker showed that the standard deviation in performance across all speakers is greatly reduced with the multimodal model.

5.2 EM Training

Although the EM algorithm is guaranteed to find a local maximum of the likelihood function for any starting point, it is not clear how different initial parameters would affect actual speaker ID performance. To investigate this issue, two initialization methods were used. In the first method, the training data was segmented into 50 pre-defined acoustic–phonetic classes, and these labeled observations were used to initialize the component Gaussian densities in Equation 1. The acoustic class means and global variance then served as the initial model for EM training. The segmentation was performed by a forced Viterbi decoding of the training utterance using 50 speaker independent hidden Markov subword models trained from read sentences. The second method consisted of randomly choosing 50 vectors from a speaker's training data for the initial model means and an identity matrix for the starting covariance matrix.

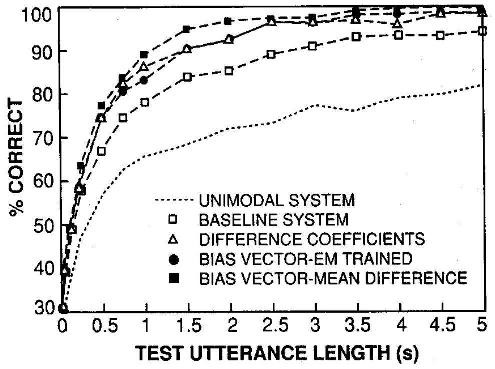

Figure 1: Speaker ID performance curves: Recognition performance curves for unimodal classifier (dotted line), baseline mixture (□), baseline with diff. coeff. (△), baseline with model based spectral norm. (•), and baseline with blind deconvolution spectral norm. (■).

Surprisingly, there was no significant difference in speaker ID performance when using either initialization method. Acoustic class segmentation showed only a 1% improvement in average identification performance for a 1 second utterance. Completely unsupervised training of speaker models certainly leads to a low complexity easily implementable system, but the issue of optimum initialization of this iterative reestimation procedure remain largely unresolved. There is a large body of work investigating the use of a priori information to improve the performance of unsupervised learning procedures, and this remains an important area of future work.

Another aspect of the EM training is the convergence rate of the algorithm. By tracking the likelihood function at each iteration while training several models, it was observed that the EM algorithm finds a maximum within 5 to 10 iterations. This convergence rate was found to be insensitive to the different initialization strategies. A preliminary experiment also demonstrated that speaker ID performance greatly suffered with less than 5 iterations but had only slight increases with more than 10 iterations. Based on these results, all models were trained with 10 iterations of the EM algorithm.

Finally, the effect of the mixture order on speaker identification performance was examined. Several versions of the baseline system with varying numbers of modes were evaluated. The performance curves for speaker models using 2, 5, 10, 15, 20, and 50 component Gaussians are shown in figure 2. From the figure, it is clear that speaker identification performance does not begin to degrade until the mixture order is less than 15.

5.3 Compensation for Intersession Speaker Variability

The last set of experiments deal with the methods for compensation of intersession variability described in section 3. For the two cepstrum bias procedures, a bias vector was computed for each baseline reference model using a test utterance and speaker identification accuracy was evaluated on the test utterance using the biased models. The performance of the biasing methods is shown in figure 1. It is interesting to see that the bias vector derived from the difference in long term means between training and test data outperformed the spectrum bias vector obtained from the EM training procedure. As a comparison to the baseline system, for a 1 second test length, recognition accuracy increased from 78.2% without compensation to 89.0% using the simple blind deconvolution approach to estimating b^ℓ.

It was also found that the use of difference cepstra resulted in a significant increase in performance. This is thought to be largely due to their relative "invariance" to a fixed linear channel component. The baseline feature vectors were augmented with 12 difference coefficients from a ±2 frame width and speaker models were trained and tested using the new vectors. As seen in figure 1, this system gives higher

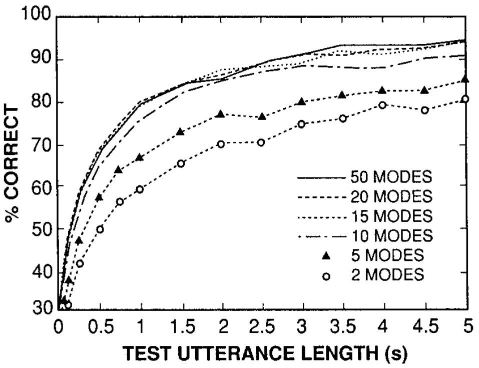

Figure 2: Performance curves for varying number of mixture components

recognition rates than the baseline system (86.3% at 1 s test length) but slightly lower rates compared to the biasing procedures. Applying the biasing methods to this system showed no significant improvement in performance.

6 DISCUSSION

A number of important observations can be made from the experimental results. The most interesting result is that consistently good speaker classification accuracy can be obtained for very short test utterances. This good performance is thought to be largely attributable to the effectiveness of the maximum likelihood Gaussian mixture in obtaining robust estimates of somewhat arbitrarily shaped underlying densities. Considerable insight can be gained by observing the individual component Gaussians represented in Equation 1 as they sum to form the complete mixture density, $p(x_k|\Theta^t)$.

Figure 3(a) shows the histogram of a cepstral feature over a 50 second utterance for a single speaker (jagged line). Superimposed over the histogram is a plot of the Gaussian density for that feature (dashed line) and a plot of the 50 component mixture density for that feature (solid line). Figure 3(b) shows each of the component Gaussians that form the mixture density plotted individually. From Figure 3(a) it is clear that the mixture model provides a better match to the tails of the true distribution than a single Gaussian model It is interesting to observe from Figure 3(b) how individual mixture modes are spread about to capture the details of the underlying density. These details probably provide important discriminant information between speakers.

The experiments also provided some insight into the parameters for the EM training. However, caution should be used when attempting to generalize experimental observations. While it was observed that initialization by acoustic class segmentation did not significantly improve performance over random initialization, it is possible that a judicious choice of a starting point is quite important for models with a smaller number of mixture components.

7 SUMMARY and CONCLUSIONS

A new method for spotting speakers in unlabeled utterances has been presented. The method is based on iterative maximum likelihood estimation of Gaussian mixtures, and provided reliable speaker identification performance over short test utterances. Mixture modeling was shown to outperform the standard unimodal Gaussian classifier over all test utterance lengths, and several methods for linear compensation were successfully applied to the mixture model ID system. A recognition rate of 89% for a 1 second test length was achieved on a 12 speaker extemporaneous speech database.

This model is currently being evaluated on a 20 speaker telephone speech database. Further work includes the investigation of discriminant techniques for training speaker models, investigation of a real–time implementation of a Gaussian mixture classifier, and the integration of Gaussian mixture based speaker spotting with speech recognition under extemporaneous speech input.

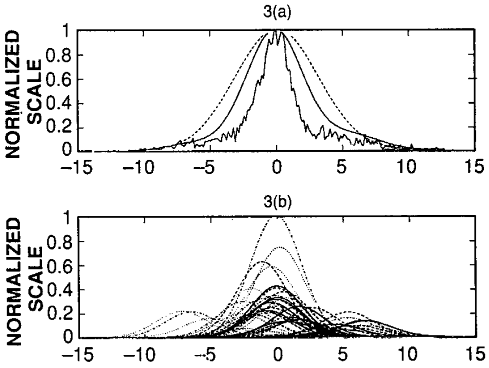

Figure 3: Example of feature distribution: (a) Histogram of feature (jagged line), best unimodal model (dashed line), and best mixture model (solid line). (b) Components of mixture model.

References

[1] S. J. Cox and J. S. Bridle. Unsupervised speaker adaptation by probabilistic spectrum fitting. *Proc. Int. Conf. on Acoust., Speech, and Sig. Processing*, May 1989.

[2] H. Gish, K. Karnofsky, M. Krasner, S. Roucos, R. Schwartz, and J. Wolf. Investigation of text-independent speaker identification over telephone channels. *Proc. Int. Conf. on Acoust., Speech, and Sig. Processing*, pages 379–382, April 1985.

[3] W. Y. Huang and R. P. Lippmann. HMM speech recognition system with neural net discrimination. *Neural Info. Processing Symposium*, November 1989.

[4] G. J. McLachlan. *Mixture Models.* Marcel Dekker, New York, N. Y., 1988.

[5] D. O'Shaunessy. Speaker recognition. *IEEE ASSP Magazine*, pages 4-17, October 1986.

[6] F. K. Soong, A. E. Rosenburg, B. H. Huang, and L. R. Rabiner. A vector quantization approach to speaker recognition. *AT&T Technical Journal*, 66(2):14-26, 1987.