Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models

Douglas A. Reynolds, *Member, IEEE,* and Richard C. Rose, *Member, IEEE*

Abstract— **This paper introduces and motivates the use of Gaussian mixture models (GMM) for robust text-independent speaker identification. The individual Gaussian components of a GMM are shown to represent some general speaker-dependent spectral shapes that are effective for modeling speaker identity. The focus of this work is on applications which require high identification rates using short utterance from unconstrained conversational speech and robustness to degradations produced by transmission over a telephone channel. A complete experimental evaluation of the Gaussian mixture speaker model is conducted on a 49 speaker, conversational telephone speech database. The experiments examine algorithmic issues (initialization, variance limiting, model order selection), spectral variability robustness techniques, large population performance, and comparisons to other speaker modeling techniques (uni-modal Gaussian, VQ codebook, tied Gaussian mixture, and radial basis functions). The Gaussian mixture speaker model attains 96.8% identification accuracy using 5 second clean speech utterances and 80.8% accuracy using 15 second telephone speech utterances with a 49 speaker population and is shown to outperform the other speaker modeling techniques on an identical 16 speaker telephone speech task.**

I. Introduction

THE speech signal conveys several levels of information. Primarily, the speech signal conveys the words or message being spoken, but on a secondary level, the signal also conveys information about the identity of the talker. While the area of speech recognition is concerned with extracting the underlying linguistic message in an utterance, the area of speaker recognition is concerned with extracting the identity of the person speaking the utterance. As speech interaction with computers becomes more pervasive in activities such as telephone financial transactions and information retrieval from speech databases, the utility of automatically recognizing a speaker based solely on vocal characteristics increases.

Depending upon the application, the general area of speaker recognition is divided into two specific tasks: verification and identification. In verification, the goal is to determine from a voice sample if a person is whom he or she claims. In speaker identification, the goal is to determine which one of a group of known voices best matches the input voice sample. Furthermore, in either task the speech can be constrained to be a known phrase (text-dependent) or totally unconstrained (text-independent). Success in both tasks depends on extracting and modeling the speaker-dependent characteristics of the speech signal which can effectively distinguish one talker from another.

In this paper a new speaker model based on Gaussian mixture models (GMM) is introduced and evaluated for text-independent speaker identification. The use of Gaussian mixture models for modeling speaker identity is motivated by the interpretation that the Gaussian components represent some general speaker-dependent spectral shapes and the capability of Gaussian mixtures to model arbitrary densities. The Gaussian mixture speaker model is experimentally evaluated on a 49 speaker conversational speech database containing both clean and telephone speech. The experiments examine algorithmic issues such as model initialization, variance limiting, and model order selection. To compensate for spectral variability introduced by the telephone channel and handsets, robustness techniques such as long-term mean removal, difference coefficients, and frequency warping are applied and compared. The experiments also examine the GMM speaker identification performance with respect to an increasing speaker population. Finally, the performance of the Gaussian mixture speaker model, uni-modal Gaussian model [1], vector quantization (VQ) codebook model [2], tied Gaussian mixture model, and radial basis function (RBF) model [3] are compared on a 16 speaker telephone speech identification task.

The techniques for speaker recognition can be categorized into three major approaches. The first and earliest approach is to use long-term averages of acoustic features, such as spectrum representations or pitch [7], [8]. The idea is to average out the other factors influencing the acoustic features, such as the phonetic variations, leaving only the speaker dependent component. For spectral features, the long-term average represents a speaker's average vocal tract shape. This approach is equivalent to a Gaussian classifier and has been used successfully for several difficult, text-independent speaker identification tasks [1], [9]. However, the averaging process discards much speaker-dependent information and can require long (>20 s) speech utterances to derive stable long-term speech statistics.

The second approach is to model the speaker-dependent acoustic features within the individual phonetic sounds that comprise the utterance. By comparing acoustic features from phonetic sounds in a test utterance with speaker-dependent acoustic features from similar phonetic sounds, the comparison measures speaker differences rather than textual difference.

Manuscript received September 8, 1993; revised May 18. 1994. This work was supported by the U.S. Department of the Air Force. The associate editor coordinating the review of this paper and approving it for publication was Dr. Joseph Campbell.

D. A. Reynolds is with the Speech Systems Technology Group, MIT Lincoln Laboratory, Lexington, MA 02173 USA.

R. C. Rose is with the Speech Research Department, AT&T Bell Laboratories, Murray Hill, NJ 07974-0636 USA.

IEEE Log Number 9406779.

This approach can be accomplished using explicit or implicit segmentation of the speech into phonetic sound classes prior to speaker model training or recognition. In [10] and [11], explicit segmentation was performed using a hidden Markov model (HMM)-based continuous speech recognizer as a front-end segmenter for text-independent speaker recognition systems. It was found in both studies that the front-end speech recognizer provided little or no improvement in speaker recognition performance compared to no front-end segmentation. Moreover, using a continuous speech recognizer front-end imposes a significant increase in computational complexity on both training and recognition.

Implicit segmentation, on the other hand, relies on some form of unsupervised clustering to provide implicit segmentation of the acoustic features during both training and recognition. The sound classes are not labeled, so separate training of a segmenter is not required. Template based clustering, such as vector quantization [12], [2] and K-nearest neighbor with leader clustering [13], has proven to be very effective for this approach to speaker recognition. In the VQ approach, each speaker is represented by a codebook of spectral templates representing the phonetic sound clusters in his/her speech. While this technique has demonstrated good performance on limited vocabulary (digits) tasks, it is limited in its ability to model the possible variabilities encountered in an unconstrained speech task. As has been shown in speech recognition, probablistic models provide a better model of acoustic speech events and a framework for dealing with noise and channel degradations. HMM's, in a variety of forms, have been used as probabilistic speaker models for both text-independent and text-dependent speaker recognition [14], [17]. The HMM models not only the underlying speech sounds, but also the temporal sequencing among these sounds. Although temporal structure modeling is advantageous for text-dependent tasks, for text-independent tasks the sequencing of sounds found in the training data does not necessarily reflect the sound sequences found in the testing data and contains little speaker-dependent information. This is supported by experimental results in [15] and [17] which found text-independent performance was unaffected by discarding transition probabilities in HMM speaker models.

The third and most recent approach to speaker recognition is the use of discriminative neural networks (NN). Rather than train individual models to represent particular speakers, discriminative NN's are trained to model the decision function which best discriminates speakers within a known set. Several different networks, such as multilayer perceptrons [18], time-delay NN's [19], and radial basis functions [3], have recently been applied to various speaker recognition tasks. Generally, NN's require a smaller number of parameters than independent speaker models and have produced good speaker recognition performance, comparable to that of VQ systems. The major drawback to many of the NN techniques is that the complete network must be retrained when a new speaker is added to the system.

The Gaussian mixture speaker model falls into the implicit segmentation approach to speaker recognition. It provides a probabilistic model of the underlying sounds of a person's voice, but unlike HMM's does not impose any Markovian constraints between the sound classes. The probabilistic framework also allows the application of newly developed noise and channel robustness techniques from the speech recognition area. In [20] a statistical background noise model is integrated with the Gaussian mixture speaker model for noise robustness using this framework. Furthermore, the new model is computationally efficient and can easily be implemented on a real-time digital signal processor [21], [22].

The research in this paper is concerned with realistic speech data encountered in practical applications of speaker identification. Speaker labeling of voice mail, for example, must use unconstrained conversational speech, possibly received over a noisy telephone line. In such an application, the speaker model must have some compensation to be robust to the acoustic distortions produced by telephone handsets and networks. Also, since there is usually no control over how long a person speaks, this research is focused on performance using short ($<$10 s) speech utterances for identification. These issues are examined in speaker identification experiments conducted on a telephone quality conversational speech database.

The rest of the paper is organized as follows. In the next section, we introduce the Gaussian mixture speaker model and motivate its use for text-independent speaker modeling. Section III then presents an experimental study of the Gaussian mixture speaker model on an unconstrained conversational database. The experiments examine parameter estimation, model order selection, spectral variability robustness, effect of population size, and performance comparisons to other speaker classifiers. Finally, Section IV gives a summary and conclusions.

II. The Gaussian Mixture Speaker Model

This section describes the form of the Gaussian mixture model (GMM) and motivates its use as a representation of speaker identity for text-independent speaker identification. The speech analysis for extracting the mel-cepstral feature representation used in this work is presented first. Next, the Gaussian mixture speaker model and its parameterization are described. The use of the Gaussian mixture density for speaker identification is then motivated by two interpretations. First, the individual component Gaussians in a speaker-dependent GMM are interpreted to represent some broad acoustic classes. These acoustic classes reflect some general speaker-dependent vocal tract configurations that are useful for modeling speaker identity. Second, a Gaussian mixture density is shown to provide a smooth approximation to the underlying long-term sample distribution of observations obtained from utterances by a given speaker. Finally, the maximum-likelihood parameter estimation and speaker identification procedures are described.

A. Speech Analysis

Although there are no exclusively speaker distinguishing speech features, the speech spectrum has been shown to be very effective for speaker identification [4]. This is because the spectrum reflects a person's vocal tract structure, the predominant physiological factor which distinguishes one person's

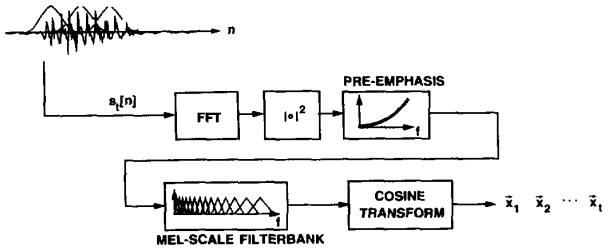


Fig. 1. Mel-scale cepstral feature analysis.

voice from others. LPC spectral representations, such as LPC cepstral and reflection coefficients, have been used extensively for speaker recognition; however, these model-based representations can be severely affected by noise [5]. Recent studies have found directly computed filterbank features to be more robust for noisy speech recognition [6]. In this paper we use cepstral coefficients derived from a mel-frequency filterbank to represent the short-time speech spectra.

Fig. 1 shows a block diagram of the steps in our front-end feature extraction. The magnitude spectrum from a 20 ms short-time segment of speech is pre-emphasized and processed by a simulated mel-scale filterbank. The filterbank follows that described in [23]. The log-energy filter outputs are then cosine transformed to produce the cepstral coefficients. The zeroth cepstral coefficient is not used in the cepstral feature vector. This processing occurs every 10 ms, producing 100 feature vectors per second.

B. Model Description

A Gaussian mixture density is a weighted sum of M component densities, as depicted in Fig. 2 and given by the equation

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \tag{1}$$

where $\vec{x}$ is a D-dimensional random vector, $b_i(\vec{x}), i = 1, \ldots, M$, are the component densities and $p_i, i = 1, \ldots, M$, are the mixture weights. Each component density is a D-variate Gaussian function of the form

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(\vec{x} - \vec{\mu}_i)' \Sigma_i^{-1} (\vec{x} - \vec{\mu}_i)\right\} \tag{2}$$

with mean vector $\vec{\mu}_i$ and covariance matrix Σ_i. The mixture weights satisfy the constraint that $\sum_{i=1}^{M} p_i = 1$.

The complete Gaussian mixture density is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. These parameters are collectively represented by the notation

$$\lambda = \{p_i, \vec{\mu}_i, \Sigma_i\} \; i = 1, \ldots, M. \tag{3}$$

For speaker identification, each speaker is represented by a GMM and is referred to by his/her model λ.

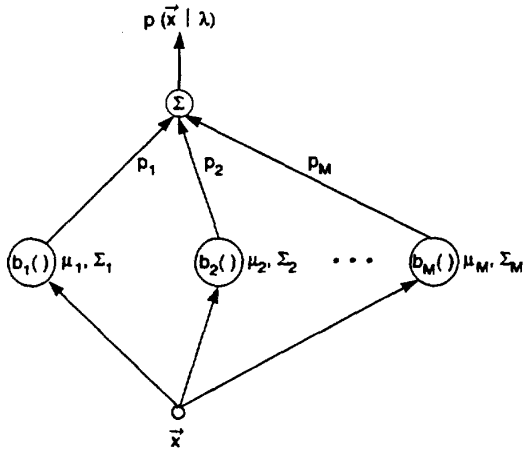


Fig. 2. Depiction of an M component Gaussian mixture density. A Gaussian mixture density is a weighted sum of Gaussian densities, where $p_i, i = 1, \ldots, M$, are the mixture weights and $b_i(), i = 1, \ldots, M$, are the component Gaussians.

The GMM can have several different forms depending on the choice of covariance matrices. The model can have one covariance matrix per Gaussian component as indicated in (3) (nodal covariance), one covariance matrix for all Gaussian components in a speaker model (grand covariance), or a single covariance matrix shared by all speaker models (global covariance). The covariance matrix can also be full or diagonal. In this paper, nodal, diagonal covariance matrices are primarily used for speaker models, except as noted for some experiments. This choice is based on initial experimental results indicating better identification performance using nodal, diagonal variances compared to nodal and grand full covariance matrices.

C. Model Interpretations

There are two principal motivations for using Gaussian mixture densities as a representation of speaker identity. The first motivation is the intuitive notion that the individual component densities of a multi-modal density, like the GMM, may model some underlying set of acoustic classes. It is reasonable to assume the acoustic space corresponding to a speaker's voice can be characterized by a set of acoustic classes representing some broad phonetic events, such as vowels, nasals, or fricatives. These acoustic classes reflect some general speaker-dependent vocal tract configurations that are useful for characterizing speaker identity. The spectral shape of the ith acoustic class can in turn be represented by the mean $\vec{\mu}_i$ of the ith component density, and variations of the average spectral shape can be represented by the covariance matrix Σ_i. Because all training or testing speech is unlabeled, the acoustic classes are "hidden" in that the class of an observation is unknown. Assuming independent feature vectors, the observation density of feature vectors drawn from these hidden acoustic classes is a Gaussian mixture.

The second motivation for using Gaussian mixture densities for speaker identification is the empirical observation that

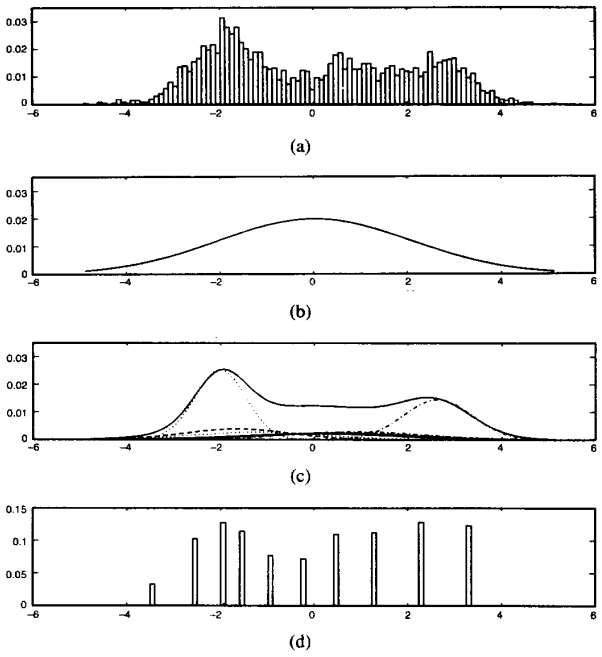


Fig. 3. Comparison of distribution modeling: (a) Histogram of a single cepstral coefficient from a 25 second utterance by a male speaker; (b) maximum likelihood unimodal Gaussian model; (c) GMM and its 10 underlying component densities; (d) histogram of the data assigned to the VQ centroid locations of a 10-element codebook.

a linear combination of Gaussian basis functions is capable of representing a large class of sample distributions. One of the powerful attributes of the GMM is its ability to form smooth approximations to arbitrarily-shaped densities. The classical unimodal Gaussian speaker model represents a speaker's feature distribution by a position (mean vector) and a elliptic shape (covariance matrix) and the VQ model represents a speaker's distribution by a discrete set of characteristic templates. In some sense the GMM acts as a hybrid between these two models by using a discrete set of Gaussian functions, each with their own mean and covariance matrix, to allow a better modeling capability. Fig. 3 compares the densities obtained using a unimodal Gaussian model, a GMM and a VQ model. Plot (a) shows the histogram of a single cepstral coefficient from a 25 second utterance by a male speaker; plot (b) shows the maximum likelihood unimodal Gaussian model; plot (c) shows the GMM and its 10 underlying component densities; and plot (d) shows a histogram of the data assigned to the VQ centroid locations of a 10-element codebook. The GMM not only provides a smooth overall distribution fit, its components also clearly detail the multi-modal nature of the density.

Also, because the component Gaussians are acting together to model the overall pdf, full covariance matrices are not necessary even if the features are not statistically independent. The linear combination of diagonal covariance Gaussians is capable of modeling the correlations between feature vector elements. The effect of using a set of M full covariance Gaussians can be equally obtained by using a larger set of diagonal covariance Gaussians.

D. Maximum Likelihood Parameter Estimation

Given training speech from a speaker, the goal of speaker model training is to estimate the parameters of the GMM, λ, which in some sense best matches the distribution of the training feature vectors. There are several techniques available for estimating the parameters of a GMM [24]. By far the most popular and well-established method is maximum likelihood (ML) estimation.

The aim of ML estimation is to find the model parameters which maximize the likelihood of the GMM, given the training data. For a sequence of T training vectors $X = \{\vec{x}_1, \ldots, \vec{x}_T\}$, the GMM likelihood can be written as

$$p(X \mid \lambda) = \prod_{t=1}^{T} p(\vec{x}_t \mid \lambda). \tag{4}$$

Unfortunately, this expression is a nonlinear function of the parameters λ and direct maximization is not possible. However, ML parameter estimates can be obtained iteratively using a special case of the expectation-maximization (EM) algorithm [25].

The basic idea of the EM algorithm is, beginning with an initial model λ, to estimate a new model $\overline{\lambda}$, such that $p(X \mid \overline{\lambda}) \geq p(X \mid \lambda)$. The new model then becomes the initial model for the next iteration and the process is repeated until some convergence threshold is reached. This is the same basic technique used for estimating HMM parameters via the Baum-Welch reestimation algorithm [26].

On each EM iteration, the following reestimation formulas are used which guarantee a monotonic increase in the model's likelihood value:

Mixture Weights:

$$\overline{p}_i = \frac{1}{T} \sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda) \tag{5}$$

Means:

$$\vec{\overline{\mu}}_i = \frac{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda) \vec{x}_t}{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda)} \tag{6}$$

Variances:

$$\overline{\sigma}_i^2 = \frac{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda) x_t^2}{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda)} - \overline{\mu}_i^2 \tag{7}$$

where σ_i^2, x_t, and μ_i refer to arbitrary elements of the vectors $\vec{\sigma}_i^{\,2}$, $\vec{x}_t$, and $\vec{\mu}_i$, respectively.

The *a posteriori* probability for acoustic class i is given by

$$p(i \mid \vec{x}_t, \lambda) = \frac{p_i b_i(\vec{x}_t)}{\sum_{k=1}^{M} p_k b_k(\vec{x}_t)}. \tag{8}$$

Two critical factors in training a Gaussian mixture speaker model are selecting the order M of the mixture and initializing the model parameters prior to the EM algorithm. There are no good theoretical means to guide one in either of these

selections, so they are best experimentally determined for a given task. An experimental examination of these factors on speaker ID performance is discussed in Section III.

E. Speaker Identification

For speaker identification, a group of S speakers $\mathcal{S} = \{1, 2, \ldots, S\}$ is represented by GMM's $\lambda_1, \lambda_2, \ldots, \lambda_S$. The objective is to find the speaker model which has the maximum *a posteriori* probability for a given observation sequence. Formally,

$$\hat{S} = \arg \max_{1 \le k \le S} \Pr(\lambda_k \mid X) = \arg \max_{1 \le k \le S} \frac{p(X \mid \lambda_k) \Pr(\lambda_k)}{p(X)} \tag{9}$$

where the second equation is due to Bayes' rule. Assuming equally likely speakers (i.e., $\Pr(\lambda_k) = 1/S$) and noting that $p(X)$ is the same for all speaker models, the classification rule simplifies to

$$\hat{S} = \arg \max_{1 \le k \le S} p(X \mid \lambda_k). \tag{10}$$

Using logarithms and the independence between observations, the speaker identification system computes

$$\hat{S} = \arg \max_{1 \le k \le S} \sum_{t=1}^{T} \log p(\vec{x}_t \mid \lambda_k) \tag{11}$$

in which $p(\vec{x}_t \mid \lambda_k)$ is given in (1).

III. Experimental Evaluation

This section presents the experimental evaluation of the Gaussian mixture speaker model for text-independent speaker identification. The GMM speaker identification system was evaluated in a task domain where utterances are from conversational speech spoken over both wideband, high signal-to-noise ratio (SNR) channels and narrowband telephone channels. The experimental study has four parts. In the first set of experiments, issues related to parameter estimation and model order selection for the Gaussian mixture speaker model are examined. The second set of experiments evaluates several different robustness techniques for improving performance using telephone speech. The third set of experiments examines the effects of speaker population size on identification performance. Finally, the last set of experiments compares the performance of the Gaussian mixture speaker model to several other classifiers, including unimodal Gaussian, vector quantization codebook, tied Gaussian mixture model, and radial basis functions.

A. Database Description

The experiments were primarily conducted using a subset of the KING speech database [27]. The KING database is a collection of conversational speech from 51 male speakers. For each speaker there are 10 conversations of approximately 45 seconds each recorded during 10 separate sessions. The speech from a session was recorded from a high-quality microphone locally and was transmitted over a long distance telephone link, providing a high-quality (clean) version and a telephone quality version of the speech. The experiments used five sessions per speaker with two-three sessions used for training data and the remaining sessions used for testing data. The model initialization experiments, described in Section III–C–1, were conducted on a different wideband, conversational speech database consisting of 12 speakers (eight male, four female).

B. Performance Evaluation

The evaluation of a speaker identification experiment was conducted in the following manner. The test speech was first processed by the front-end analysis to produce a sequence of feature vectors $\{\vec{x}_1, \ldots, \vec{x}_t\}$[1]. To evaluate different test utterance lengths, the sequence of feature vectors was divided into overlapping segments of T feature vectors. The first two segments from a sequence would be

$$\overbrace{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_T}^{\text{Segment 1}}, \vec{x}_{T+1}, \vec{x}_{T+2}, \ldots$$

$$\vec{x}_1, \overbrace{\vec{x}_2, \ldots, \vec{x}_T, \vec{x}_{T+1}}^{\text{Segment 2}}, \vec{x}_{T+2}, \ldots.$$

A test segment length of 5 seconds corresponds to $T = 500$ feature vectors at a 10 ms frame rate. Each segment of T vectors was treated as a separate test utterance.

The identified speaker of each segment was compared to the actual speaker of the test utterance and the number of segments which were correctly identified was tabulated. The above steps were repeated for test utterances from each speaker in the population. The final performance evaluation was then computed as the percent of correctly identified T-length segments over all test utterances

$$\%\ \text{correct identification} = \frac{\#\ \text{correctly identified segments}}{\text{total}\ \#\ \text{of segments}} \times 100. \tag{12}$$

The evaluation was repeated for different values of T to evaluate performance with respect to test utterance length.

Each speaker had approximately equal amounts of testing speech so the performance evaluation was not biased to any particular speaker. While there may be variations among the individual speakers' performance, the aim of the evaluation measure was to track the average performance of the system for different speaker identification tasks, allowing a common basis of comparison.

C. Algorithmic Issues

1) Initialization: As stated in the previous section, the GMM training procedure must be initialized with some starting model $\lambda^{(0)}$. The EM algorithm is guaranteed to find a local maximum likelihood model regardless of the starting point, but the likelihood equation for a GMM has several local maxima and different starting models can lead to different local maxima [24]. To investigate the effect of model initialization on speaker identification performance, speaker models were

[1]Periods of silence are removed from the test speech prior to feature extraction using an adaptive energy threshold speech/silence detector.

trained using different methods of initialization and used for a speaker identification experiment. This experiment used the 12 speaker conversational database. Speakers were modeled by a 50 component GMM with a grand, diagonal covariance matrix trained using approximately 5000 12-dimensional mel-cepstral vectors (50 seconds). Testing was done using approximately three minutes of speech per speaker.

The first method of initialization used a speaker-independent HMM to automatically segment the training speech. The training data was segmented into 50 labeled phonetic classes which corresponded to the initial mixture components. The class means and global variances then served as the initial model for EM training. The segmentation was performed by a forced Viterbi decoding of the unlabeled training utterance using monophone acoustic models. The acoustic models were obtained from averaging speaker-independent, context-dependent subword HMM's. The subword HMM's were trained using the forward-backward algorithm on orthographically transcribed continuous speech utterances. The second initialization method consisted of randomly choosing 50 vectors from a speaker's training data (after silence removal) for the initial model means and an identity matrix for the starting covariance matrix.

Surprisingly, no significant difference in speaker identification performance was found between the two initialization methods. The different initial models may have converged to different local maximizers of the likelihood function, but the difference between the final models is insignificant in terms of speaker identification performance. It was also observed that both methods of initialization required the same number of EM iterations for convergence of the likelihood function, so no training speed advantage was found for either method. These results indicate that elaborate initialization schemes are not necessary for training Gaussian mixture speaker models.

Subsequent experiments also found no significant difference between the above random mean selection and binary k-means clustering for initialization. The rest of the experiments in this paper use random mean selection, followed by a single iteration k-means clustering to initialize means, nodal variances, and mixture weights.

2) Variance Limiting: When training a nodal variance GMM, it has been observed that variance elements can become quite small in magnitude. This is particularly true for a mixture model with a large ($\geq$32) number of component densities. These small variances produce a singularity in the model's likelihood function and can degrade identification performance by distorting speaker model scores used in the maximum likelihood classifier. These singularities can arise when there is not enough data to sufficiently train a component's variance vector or when using noise-corrupted data. The noisy data can contain outliers in the data that give rise to components with very small variances [28].

To avoid these spurious singularities, a variance limiting constraint is applied. This constraint places a minimum variance value on elements of all variance vectors in a speaker's model. For an arbitrary element of mixture component i's variance vector, σ_i^2, and a minimum variance value, $\sigma_{\min}^2$, the constraint

$$\overline{\sigma}_i^2 = \begin{cases} \sigma_i^2 & \text{if } \sigma_i^2 > \sigma_{\min}^2 \\ \sigma_{\min}^2 & \text{if } \sigma_i^2 \leq \sigma_{\min}^2 \end{cases} \tag{13}$$

is applied to the variance estimates after each EM iteration to avoid singularities in the final model. This is a constrained version of the EM algorithm which has been been shown to provide more robust parameter estimates than the unconstrained version [24], [29].

Care must be exercised when setting the minimum variance value. If it is set too high, the component variances are masked to the same value which would overly constrain the model and hence degrade identification performance. Setting the value too low may not perform the desired limiting at all. The variance limit must be empirically determined for any particular data set, feature set, and model size to optimize performance. Preliminary experiments on a 16 speaker set found a variance limit between $\sigma_{\min}^2 = 0.01$ and $\sigma_{\min}^2 = 0.1$ to provide the best robustness for mel-cepstral features.

3) Model Order: Determining the number of components M in a mixture needed to model a speaker adequately is an important but difficult problem. There is no theoretical way to estimate the number of mixture components *a priori*. For speaker modeling the objective is to choose the minimum number of components necessary to adequately model a speaker for good speaker identification. Choosing too few mixture components can produce a speaker model which does not accurately model the distinguishing characteristics of a speaker's distribution. Choosing too many components can reduce performance when there are a large number of model parameters relative to the available training data and can also result in excessive computational complexity both in training and classification. The following experiments examine the performance of the GMM speaker ID system for different model orders using a fixed and variable amount of training data.

To investigate the speaker identification performance of the GMM with respect to the number of component densities per model, the following experiment was conducted on a 16 speaker subset of the KING database. Speaker models with 1, 2, 4, 8, 16, 32, and 64 component Gaussian densities and nodal variances were trained using 6000 25-dimensional mel-cepstral vectors corresponding to one minute of speech. Sessions one and two were used for model training and sessions three, four, and five were used for testing. Variance limiting was used with $\sigma_{\min}^2 = 0.01$. Fig. 4 shows the percent correct identification performance versus the number of Gaussian components for 1, 5, and 10 second test utterance lengths.

There are several observations to be made from these results. First, the sharp increase in identification performance from 1 to 8 mixture components, and leveling off above 16 components, indicates that there is a lower limit on the number of mixture components necessary to adequately model the speakers. Models must contain at least this minimum number of components to maintain good speaker identification performance. This limit seems to be 16 mixture components for these speakers and this

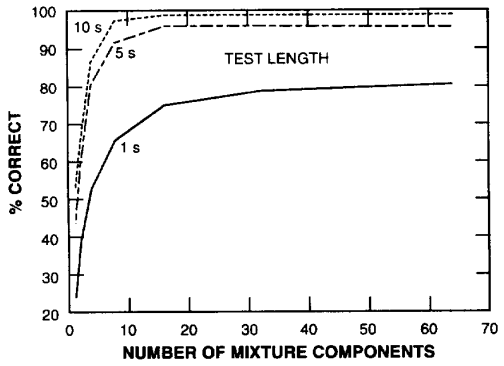


Fig. 4. Speaker identification performance as a function of the number of component densities per speaker model.

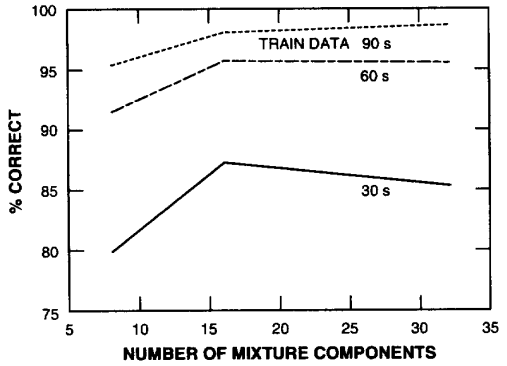


Fig. 5. Speaker identification performance versus model order for models trained with 30, 60 and 90 seconds of speech. The test utterance length is 5 seconds.

TABLE I
GMM IDENTIFICATION PERFORMANCE FOR DIFFERENT AMOUNTS OF TRAINING DATA AND MODEL ORDERS

Amount of Training Speech	Model Order	Test Length		
		1 sec	5 sec	10 sec
30 sec	$M = 8$	54.6	79.8	85.6
	$M = 16$	63.7	87.3	90.5
	$M = 32$	64.6	85.3	88.4
60 sec	$M = 8$	66.1	91.5	97.3
	$M = 16$	74.9	95.7	98.8
	$M = 32$	78.6	95.6	98.3
90 sec	$M = 8$	71.5	95.5	98.8
	$M = 16$	79.0	98.0	99.7
	$M = 32$	84.7	98.8	99.6

data. Above this minimum model order, the performance is insensitive to the number of mixture components for the 5 and 10 second length test utterances. For the 1 second test utterance length, the identification performance continues to increase (at a decreasing rate) with the model order. This demonstrates how additional components, which model additional acoustic classes, are effectively used for short utterance identification. The increase in performance begins to level out above 32 component Gaussians.

In the next experiment, speaker models with 8, 16, and 32 component densities were trained using 30, 60, and 90 seconds of speech in the same manner as described above. The various amounts of training data were sequentially taken from sessions one, two, and three and sessions four and five were used for testing. Table I shows the complete identification results. For each model order, the identification performance for 1, 5, and 10 second test utterance lengths are given for 30, 60, and 90 seconds of training data.

As expected, with increased training data, identification performance increases. Identification rates for shorter test utterance lengths show the greatest improvement. The largest increase in performance occurs at all test utterance lengths when the amount of training data increases from 30 to 60 seconds. Increasing the training data to 90 seconds also improves performance but with a smaller increment. This suggests that at least one minute of conversational speech is necessary to maintain high speaker identification performance and using more training data improves performance at a decreasing rate. Note, however, that each increase in the training data also adds training data from another session. Thus, the addition of data from different sessions may also be a factor, along with the increase in amount of data.

It is also evident that model order selection becomes more important with smaller amounts of training data. Fig. 5 plots the identification performance for the 5-second test utterance length versus model order for models trained with different amounts of training data. For all amounts of training data, performance peaks at 16 components. However, performance decreases for the 32 mixture model trained with only 30 seconds of speech. Compared with the constant or slightly increasing performance using 60 and 90 seconds of speech, this is a good example of the effects of having insufficient training samples relative to the number of model parameters being estimated.

D. Spectral Variability Compensation

The major spectral degradation found in speech collected from the telephone network is a filtering effect which band limits and imposes some spectral shaping on the speech spectrum [30]. Left uncompensated, this degradation can produce severe reductions in identification performance due to data mismatch between training and recognition data. As a first-order model, the spectral variability introduced by a telephone channel can be modeled by a linear filter effect which modifies the spectral features used by the GMM speaker ID system. Below, some spectral variability compensation techniques to produce robust features for telephone quality speech are described.

1) Frequency Warping: To avoid any differences in channel bandwidth and using any spurious out-of-band spectral components, frequency warping was applied to the magnitude DFT spectrum. The warping maps the frequency axis f to a new frequency axis f' according to the equation

$$f' = \frac{f - f_{\min}}{f_{\max} - f_{\min}} f_N \tag{14}$$

where f_N is the original Nyquist frequency. The linear warping both eliminates spectral components outside the specified

frequency range $[f_{\min}, f_{\max}]$ and expands the spectrum to full bandwidth for subsequent processing.

2) Spectral Shape Compensation: When the speech signal passes through a linear filter $h[n]$ representing the telephone channel, its magnitude spectrum is multiplied by the magnitude response of the filter. If it is assumed that the magnitude spectrum of the filter is relatively smooth, it can be shown that the effect of the filter is an additive component on the mel-cepstral feature vector [31]

$$\vec{z} = \vec{x} + \vec{h} \tag{15}$$

where $\vec{z}$ is the observed cepstral vector, $\vec{h}$ is the channel filter cepstral vector, and $\vec{x}$ is the input speech cepstral vector.

The aim of the spectral shape compensation is to remove the "bias" term $\vec{h}$ from the feature vectors. Two methods were applied to the GMM speaker identification system: mean normalization and time difference coefficients.

The method of mean normalization has been used in many speaker recognition systems [32]–[35]. Essentially, it consists of removing the bias component by subtracting off the global average vector from each feature vector. For a set of feature vectors $\{\vec{z}_t\}_{t=1}^{T}$, the global average vector is

$$\vec{m} = \frac{1}{T}\sum_{t=1}^{T}\vec{z}_t \tag{16}$$

and the channel compensated vectors are given by

$$\vec{z}_t^{\,\text{comp}} = \vec{z}_t - \vec{m}. \tag{17}$$

For each channel from which speech was collected, the global mean is subtracted off each vector before training a speaker model or scoring for recognition. All feature vectors then have the same global mean and speaker discrimination is not affected by different channel biases.

The above assumes a time-invariant channel filter. If the channel filter is time-varying, an adaptive bias removal method, such as RASTA processing [36], can be used to remove the time-varying channel bias.

Besides removing the channel filter bias, this compensation also removes the global mean of the speech feature vectors. This is equivalent to filtering the speech by the inverse of its average spectrum. Although the average speech spectrum does contain speaker specific information, it exhibits significant intra-speaker variability over time [37], [35] which can decrease recognition performance when training and testing on speech collected at different times. The average spectra is also susceptible to variations due to speech effort (for example, loud or soft) and health (for example, a speaker has a cold). Using mean normalization for clean speech improves identification performance by minimizing intersession variability. When used on telephone speech, removal of the global average minimizes both intersession variability and removes the spectral shaping imposed by different telephone channels.

Another way to minimize the channel filter effects is to use "channel invariant" features. One such set of channel invariant features used in speaker recognition systems is cepstrum difference coefficients [38]–[40].

The motivation in using difference coefficients is both to capture dynamic information and to remove time-invariant spectral information generally attributed to the interposed communications channel. This is accomplished by creating a new set of features as the time difference between the cepstral feature vectors. For frame t the difference coefficients, denoted $\Delta\vec{z}_t$, are formed by taking the difference between cepstral feature vectors that are W frames apart:

$$\Delta\vec{z}_t = \vec{z}_t - \vec{z}_{t-W}. \tag{18}$$

Since the channel filter is time-invariant or slowly varying, the bias term $\vec{h}$ in (15) is removed, leaving the difference in speech cepstra

$$\Delta\vec{z}_t = \vec{x}_t - \vec{x}_{t-W}. \tag{19}$$

Because the difference coefficients capture the spectral changes in time, they are also referred to as transitional or dynamic features, with the cepstral vectors called instantaneous or static features.

Difference coefficients have been shown to contain speaker specific information and to be fairly uncorrelated with the static cepstral feature vectors; however, when used by themselves, they do not perform as well as the static feature vectors [39]. To combine the two feature sets, the difference coefficients are appended to the cepstral feature vectors. The new feature vector not only contains channel invariant features but also spectral transitional information along with the instantaneous cepstral coefficients.

Using the above compensation techniques, speaker identification experiments were conducted using speech from different telephone channels. The experiments were conducted using the telephone version speech sessions from a 16 speaker subset of the KING database. Each speaker was modeled by a 50 component GMM trained with speech from sessions one, two, and three (which corresponds to an average of 80 seconds of training speech) using 20-dimensional mel-cepstrum feature vectors and a variance limit of $\sigma_{\min}^2 = 0.1$. The experiment with difference coefficients used 20 cepstral coefficients appended with 20 difference coefficients from a 40 ms interval (± 2 frames) around the current frame. For frequency warping, the telephone bandwidth 300–3300 Hz was linearly warped to full bandwidth. The identification results using the different compensation techniques are shown in Fig. 6.

It is evident that without compensation, speaker identification performance was degraded using telephone speech. For a similar experiment using the clean speech versions of the sessions, a speaker identification accuracy of 94.3% was attained for a 5 second test utterance compared to 64.4% using the uncompensated telephone speech. Of the compensation techniques, the most effective method for minimizing the channel variation effects was mean normalization. This simple method improved the performance by an average 28% over all test utterance lengths. Spectrum frequency warping alone produced a substantial 21% increase compared to no compensation and was the second best effective method of compensation. Using appended difference coefficients provided a

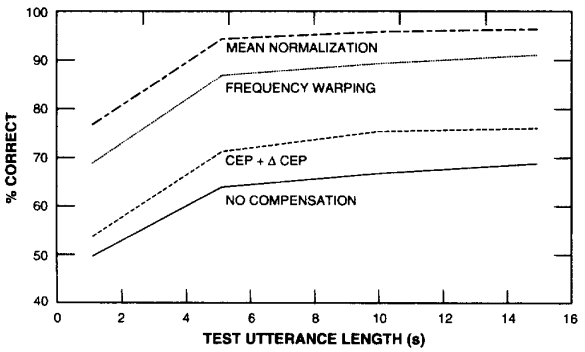


Fig. 6. Identification performance for different spectral variability compensation techniques applied to telephone speech.

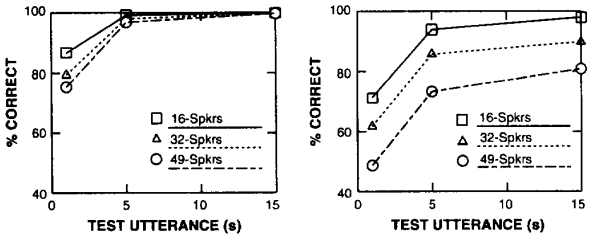


Fig. 7. Speaker identification performance versus test utterance length for population sizes of 16, 32, and 49 speakers: (a) Clean speech performance; (b) telephone speech performance.

modest improvement performance over no compensation on the order of 6%. Performing frequency warping prior to mean normalization and using mean normalized features appended with difference coefficient showed no significant improvement over mean normalization alone.

E. Large Population Performance

One factor which defines the difficulty of the speaker identification task is the size of the speaker population. As the number of speakers that the system must distinguish increases, the probability of an incorrect classification increases. The similarity of the speakers in the population also must be considered, since a set of speakers with dissimilar voice characteristics (e.g., a population of half males and half females) generally produces higher identification performance than a more homogeneous set of speakers (e.g., all male). The following experiments examined the performance of the GMM speaker identification system as a function of population size for an all-male collection of speakers using both clean and telephone speech.

For these experiments, each speaker was modeled by a 50 component GMM with nodal variances using 20-dimensional mel-cepstral feature vectors. The models were trained using all the data from sessions one, two, and three (80–100 seconds of speech per speaker) and testing was conducted using sessions four and five. Each session was mean normalized to minimize inter-session variability and channel bias. A variance limit of $\sigma^2_{\min} = 0.1$ was used in training. For the telephone speech, the frequency bandwidth 300–3300 Hz was warped to full bandwidth.

Identification performance versus test utterance lengths for populations of 16, 32, and 49 speakers are shown in Fig. 7. In the clean speech case, it is clear that the GMM speaker ID system maintains high identification performance as the population size increases. The largest degradation for increasing population size is for the 1-second test utterance length, but almost perfect identification for all population sizes is obtained for 15 second test utterances.

When compared to the clean speech results, there was a marked decrease in performance using telephone speech. One major contributing factor to this reduced performance is the relatively low SNR of some of the telephone speech. The compensation techniques only address spectral variability and so significant differences in noise levels between channels were not compensated. Examination of the telephone speech found that half of the speakers' telephone speech sessions are very noisy (SNR ranges roughly from 10 to 20 dB). The 16 speaker population used moderate SNR (approximately 30 dB) speech and the results are comparable to those using clean speech. However, as the population size increased, more speakers with noisy speech were added to the population and the performance rapidly declined.

F. Comparison to Other Speaker Models

The last set of experiments compared the performance of the Gaussian mixture speaker model with other speaker modeling techniques. Specifically, the other techniques are the unimodal Gaussian classifier (GC) [1], vector quantization (VQ) codebook [2], tied Gaussian mixture model (TGMM) and radial basis function (RBF) [3]. The aim is to compare the performance of these different identification methods using the same data and front-end processing.

These different speaker modeling techniques are interesting to compare because they represent different ways of modeling the speaker's acoustic feature distribution. In the simplest case, the GC models each speaker's feature distributions by a unimodal Gaussian distribution. Since the data is mean normalized, the Gaussian mean vector is effectively zero and identification is based only on the covariance modeling of the data. This is similar to "covariance-only" speaker identification method in [9]. The VQ models the distributions by representative templates from hard partitions of the feature space. As discussed earlier, the GMM generalizes on this notion by providing a soft partitioning of a speaker's space using Gaussian basis functions.

The RBF and TGMM both share the same underlying structure as the GMM, but model the feature space in different ways (see Fig. 8). The TGMM uses a pool of Gaussians which covers the feature space of all speakers. A maximum likelihood training procedure adjusts each speaker's mixture weights to the underlying Gaussians to best model his/her feature distribution. The underlying Gaussians' parameters are also updated in the training to match the overall feature distribution. The RBF differs from the above models in that it focuses on modeling the boundary regions separating speaker distributions in the feature space. Like the TGMM, it too uses

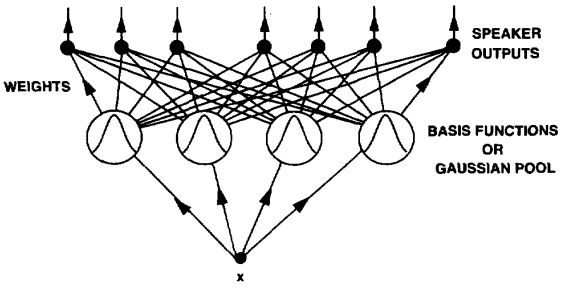


Fig. 8. TGMM and RBF model structure. In each model, speakers are represented by a weighted combination of a common pool of Gaussian or basis functions.

a pool of basis functions to represent all speakers. However, the basis functions are fixed during training and the speaker's connection weights are trained using a discriminative criterion.

The data used for the experiment was from the 16 speaker KING subset using the telephone speech sessions. All sessions were mean normalized prior to training and testing. Each model was trained using all the speech from sessions one, two, and three, with testing performed on sessions four and five. Twenty-dimensional mel-cepstral feature vectors were used and trained variances were limited to $\sigma^2_{\min} = 0.1$.

Model parameters were set as follows. Two forms of the GMM were used where the first form (GMM-nv) had 50 components with nodal variances and the second form (GMM-gv) also had 50 components, but with a single grand variance per model. The VQ-50 speaker model used 50 vectors per codebook while the VQ-100 model used 100 vectors per codebook, both trained with the LBG algorithm [41] using the Mahalanobis distance with a global, diagonal covariance matrix. The tied Gaussian mixture model used a pool of 800 Gaussians and a global diagonal covariance matrix. The radial basis function used 512 basis functions with empirically determined function widths. Finally, the unimodal Gaussian classifier used a full 20×20 covariance matrix.

The average number of parameters per speaker for each model is shown in Table II. For example, the number of parameters for the GMM-gv is calculated as (#mean_vecs + #variance_vecs) $\times$ vec_dim + #mixture_weights = (50+1) $\times$ 20+50=1070. The GMM-nv has the most parameters due to the use of nodal variances, while the GC has the least number of parameters due to the limited model structure. The GMM-nv and VQ-100 models have comparable number of parameters, as do the GMM-gv and VQ-50 models. The TGMM and RBF have different number of parameters because numerical difficulties prevented the training of an RBF with 800 basis functions.

Table III shows the percent correct identification for 5 second test utterance lengths for the different models. Also shown is the binomial standard deviation of the tests using only the number of nonoverlapping test intervals as the number of trials ($n = 160$). The classifiers can be divided into four levels of performance. On the top level, the nodal variance GMM (GMM-nv) has the best absolute performance with the VQ-100 about 1.5 percentage points lower. On the second level, the grand variance GMM (GMM-gv), VQ-50, and RBF all have similar classification performances. The drop in performance of the GMM going from nodal to grand variances indicates the importance of variance parameterization in model selection. Also, note that although the RBF has fewer centers per speaker compared to the GMM-gv and VQ models, it maintains similar performance due to the discriminative training. On the third level, the TGMM has significantly lower classification performance. This is likely due to the overly-restrictive constraint of using a single global variance vector. Lastly, the GC, using only covariance matrices, produced the worst identification performance of the classifiers.

TABLE II
NUMBER OF PARAMETERS PER SPEAKER FOR SPEAKER MODELS DISCUSSED IN TEXT

Speaker Model	Avg number of parameters per speaker
GMM-nv	2050
VQ-100	2001
GMM-gv	1070
VQ-50	1001
RBF	1152
TGMM	1801
GC	210

TABLE III
SPEAKER IDENTIFICATION PERFORMANCE FOR SPEAKER MODELS DISCUSSED IN TEXT

Speaker Model	% Correct Identification (5 second test length)
GMM-nv	94.5 ±1.8
VQ-100	92.9 ±2.0
GMM-gv	89.5 ±2.4
VQ-50	90.7 ±2.3
RBF	87.2 ±2.6
TGMM	80.1 ±3.1
GC	67.1 ±3.7

IV. CONCLUSION

This paper has introduced and evaluated the use of Gaussian mixture speaker models for robust text-independent speaker identification. The primary focus of this work was on a task domain for real applications, such as voice mail labeling and retrieval. The Gaussian mixture speaker model was specifically evaluated for identification tasks using short duration utterances from unconstrained conversational speech, possibly transmitted over noisy telephone channels.

Gaussian mixture models were motivated for modeling speaker identity based on two interpretations. The component Gaussians were first shown to represent characteristic spectral shapes (vocal tract configurations) from the phonetic sounds which comprise a person's voice. By modeling the underlying acoustic classes, the speaker model is better able to model the short-term variations of a person's voice, allowing high identification performance for short utterances. The Gaussian mixture speaker model was also interpreted as a nonparametric, multivariate pdf model, capable of modeling arbitrary feature distributions.

The experimental evaluation examined several aspects of using Gaussian mixture speaker models for text-independent speaker identification. Some observations and conclusions are:

- Identification performance of the Gaussian mixture speaker model is insensitive to the method of model initialization.
- Variance limiting is important in training to avoid model singularities.
- There appears to be a minimum model order needed to adequately model speakers and achieve good identification performance (Sixteen for this 16 speaker database).
- The Gaussian mixture speaker model maintains high identification performance with increasing population size (The system attained a 96.8% identification rate for 5 second clean speech utterances and 80.8% for 15 second telephone speech utterances for an all-male 49 speaker population).
- Cepstral mean normalization is a very effective compensation for telephone spectral variability degradations.
- With nodal variance parameterization, the Gaussian mixture speaker model outperforms the VQ, RBF, TGMM, and GC speaker modeling techniques on an identical telephone speech task.

These results indicate that Gaussian mixture models provide a robust speaker representation for the difficult task of speaker identification using corrupted, unconstrained speech. The models are computationally inexpensive and easily implemented on a real-time platform [21], [22]. Furthermore, their probabilistic framework allows direct integration with speech recognition systems [42] and incorporation of newly developed speech robustness techniques [20].

ACKNOWLEDGMENT

The authors wish to thank Capt. M. S. Ciancetta for his help on the large population and speaker model comparison experiments.

REFERENCES

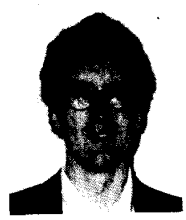
[1] H. Gish *et al.*, "Investigation of text-independent speaker identification over telephone channels," in *Proc. IEEE ICASSP*, 1985, pp. 379–382.
[2] F. Soong *et al.*, "A vector quantization approach to speaker recognition," in *Proc. IEEE ICASSP*, 1985, pp. 387–390.
[3] J. Oglesby and J. Mason, "Radial basis function networks for speaker recognition," in *Proc. IEEE ICASSP*, May 1991, pp. 393–396.
[4] B. Atal, "Automatic recognition of speakers from their voices," *Proc. IEEE*, vol. 64, pp. 460–475, Apr. 1976.
[5] J. Tierney, "A study of LPC analysis of speech in additive noise," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 389–397, Aug. 1980.
[6] C. R. Jankowski, unpublished research, MIT Lincoln Laboratory.
[7] S. Furui, F. Itakura, and S. Saito, "Talker recognition by longtime averaged speech spectrum," *Electron., Commun. in Japan*, vol. 55-A, no. 10, pp. 54–61, 1972.
[8] J. Markel, B. Oshika, and A. Gray, Jr., "Long-term feature averaging for speaker recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 330–337, Aug. 1977.
[9] H. Gish *et al.*, "Methods and experiments for text-independent speaker recognition over telephone channels," in *Proc. IEEE ICASSP*, 1986, pp. 865–868.
[10] T. Matsui and S. Furui, "A text-independent speaker recognition method robust against utterance variations," in *Proc. IEEE ICASSP*, 1991, pp. 377–380.
[11] Y. Kao, P. Rajasekaran, and J. Baras, "Free-text speaker identification over long distance telephone channel using hypothesized phonetic segmentation," in *Proc. IEEE ICASSP*, 1992, pp. II. 177–II. 180.
[12] R. E. Helms, "Speaker recognition using linear predictive vector codebooks," Ph.D. thesis, Southern Methodist University, 1981.
[13] A. Higgins, L. Bahler, and J. Porter, "Voice identification using nearest-neighbor distance measure," in *Proc. IEEE ICASSP*, Apr. 1993, pp. II-375–II-378.
[14] A. B. Poritz, "Linear predictive hidden Markov models and the speech signal," in *Proc. IEEE ICASSP*, May 1982, pp. 1291–1294.
[15] N. Z. Tishby, "On the application of mixture AR hidden Markov models to text independent speaker recognition," *IEEE Trans. Signal Processing*, vol. 39, pp. 563–570, Mar. 1991.
[16] A. E. Rosenberg, C. H. Lee, and F. K. Soong, "Sub-word talker verification using hidden Markov models," in *IEEE ICASSP*, Apr. 1990, pp. 269–272.
[17] T. Matsui and S. Furui, "Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs," in *Proc. IEEE ICASSP*, Mar. 1992, pp. II.157–II.164.
[18] L. Rudasi and S. A. Zahorian, "Text-independent talker identification with neural networks," in *Proc. IEEE ICASSP*, May 1991, pp. 389–392.
[19] Y. Bennani and P. Gallinari, "On the use of TDNN-extracted features information in talker identification," in *Proc. IEEE ICASSP*, May 1991, pp. 385–388.
[20] R. C. Rose, E. M. Hofstetter, and D. A. Reynolds, "Integrated models of speech and background with application to speaker identification in noise," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 2, pp. 245–257, Apr. 1994.
[21] D. A. Reynolds, "A Gaussian mixture modeling approach to text-independent speaker identification," Ph.D. thesis, Georgia Inst. of Technology, Sept. 1992.
[22] D. A. Reynolds, R. C. Rose, and M. J. T. Smith, "PC-based TMS320C30 implementation of the Gaussian mixture model text-independent speaker recognition system," in *Proc. Int. Conf. Signal Processing Appl., Technol.*, Nov. 1992, pp. 967–973.
[23] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 357–366, Aug. 1980.
[24] G. McLachlan, *Mixture Models*. New York: Marcel Dekker, 1988.
[25] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Stat. Soc.*, vol. 39, pp. 1–38, 1977.
[26] L. Baum *et al.*, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math Stat.*, vol. 41, pp. 164–171, 1970.
[27] J. Godfrey, D. Graff, and A. Martin, "Public databases for speaker recognition and verification," in *Proc. ESCA Workshop Automat. Speaker Recognition, Identification, Verification*, Apr. 1994, pp. 39–42.
[28] J. Holmes and N. Sedgwick, "Noise compensation for speech recognition using probabilistic models," in *Proc. IEEE ICASSP*, 1986.
[29] R. Hathaway, "A constrained formulation of maximum-likelihood estimation for normal mixture distributions," *Ann. Stat.*, vol. 13, no. 2, pp. 795–800, 1985.
[30] J. G. Proakis, *Digital Communications*. New York: McGraw-Hill Series in Electrical Engineering, 1983.
[31] D. A. Reynolds and R. C. Rose, "An integrated speech-background model for robust speaker identification," in *Proc. IEEE ICASSP*, Mar. 1992, pp. II-185–II-188.
[32] B. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, vol. 55, pp. 1304–1312, June 1974.
[33] S. Furui, "Comparison of speaker recognition methods using statistical features and dynamic features," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 342–350, June 1981.
[34] M. Krasner *et al.*, "Investigation of text-independent speaker identification techniques under conditions of variable data," in *Proc. IEEE ICASSP*, 1984, pp. 18B.5.1–4.
[35] C. Bernasconi, "On instantaneous and transitional spectral information for text-dependent speaker verification," *Speech Commun.*, vol. 9, pp. 129–139, Apr. 1990.
[36] H. Hermansky *et al.*,"RASTA-PLP speech analysis technique," in *Proc. IEEE ICASSP*, Mar. 1992, pp. I.121–I.124.
[37] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 254–272, Apr. 1981.
[38] R. E. Bogner, "On talker verification via orthogonal parameters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 1–12, Feb. 1981.

[39] F. Soong and A. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 871–879, June 1988.
[40] R. C. Rose and D. A. Reynolds, "Text-independent speaker identification using automatic acoustic segmentation," in *Proc. IEEE ICASSP*, 1990, pp. 293–296.
[41] R. Gray, "Vector quantization," *IEEE ASSP Magazine*, pp. 4–29, Apr. 1984.
[42] D. A. Reynolds and L. P. Heck, "Integration of speaker and speech recognition systems," in *Proc. IEEE ICASSP*, May 1991, pp. 869–872.

Douglas A. Reynolds (S'86–M'92) received the B.E.E. degree with highest honors in 1986 and the Ph.D. degree in electrical engineering in 1992, both from the Georgia Institute of Technology.

Currently, he is a Staff Member in the Speech Systems Technology Group at the Massachusetts Institute of Technology Lincoln Laboratory, where his research interests include robust speaker identification and verification, speech recognition, and transient signal classification.

Dr. Reynolds is a member of Eta Kappa Nu and Tau Beta Pi.



Richard C. Rose (S'86–M'88) received the B.S. and M.S. degrees in electrical engineering from the University of Illinois in 1979 and 1981, respectively. He received the Ph.D. degree in electrical engineering from the Georgia Institute of Technology in 1988, completing his dissertation work in speech coding and analysis.

From 1980 to 1984, he was with Bell Laboratories, Holmdel, NJ, where he worked on speech processing problems in digital switching environments. From 1988 to 1992, he was a member of the Speech Systems Technology Group at the MIT Lincoln Laboratory. While there, he was involved in developing techniques for keyword recognition, improved noise robustness in speech processing, and speaker identification. He is presently a member of the technical staff at Bell Laboratories, Murray Hill, NJ, where his work has focused on problems relating to speech recognition and speaker verification.

Dr. Rose is a member of the IEEE Signal Processing Society Technical Committee on Digital Signal Processing and the Acoustical Society of America Technical Committee on Speech. He is an adjunct faculty member with the Georgia Institute of Technology. He is a member of Tau Beta Pi, Eta Kappa Nu, and Phi Kappa Phi.