ELSEVIER

Speech Communication 17 (1995) 91–108

SPEECH COMMUNICATION

Speaker identification and verification using Gaussian mixture speaker models ☆

Douglas A. Reynolds *

MIT Lincoln Laboratory, 244 Wood St, Lexington, MA 02173, USA

Received 27 September 1994; revised 9 March 1995

Abstract

This paper presents high performance speaker identification and verification systems based on Gaussian mixture speaker models: robust, statistically based representations of speaker identity. The identification system is a maximum likelihood classifier and the verification system is a likelihood ratio hypothesis tester using background speaker normalization. The systems are evaluated on four publically available speech databases: TIMIT, NTIMIT, Switchboard and YOHO. The different levels of degradations and variabilities found in these databases allow the examination of system performance for different task domains. Constraints on the speech range from vocabulary-dependent to extemporaneous and speech quality varies from near-ideal, clean speech to noisy, telephone speech. Closed set identification accuracies on the 630 speaker TIMIT and NTIMIT databases were 99.5% and 60.7%, respectively. On a 113 speaker population from the Switchboard database the identification accuracy was 82.8%. Global threshold equal error rates of 0.24%, 7.19%, 5.15% and 0.51% were obtained in verification experiments on the TIMIT, NTIMIT, Switchboard and YOHO databases, respectively.

Zusammenfassung

Dieses Referat befaßt sich mit Hochleistungssystemen zur Sprechererkennung und Sprecherverifizierung auf der Basis von normalverteilten Sprechermodellen, d.h. robusten, statistisch ausgewogenen Repräsentationen der Sprecheridentität. Bei dem Erkennungssystem handelt es sich um einen Klassierer nach dem Maximum-Likelihood-Prinzip; das Verifikationssystem ist ein Likelihoodverhältnis-Hypothesentester mit Hintergrund-Normalisierung für die Sprechmuster. Die Bewertung der Systeme erfolgt anhand von vier öffentlich zugänglichen Sprachdatenbanken (TIMIT, NTIMIT, Switchboard und YOHO). Die bei den Sprachmustern in diesen Datenbanken bestehenden unterschiedlichen Qualitätsverluste und Schwankungen lassen die Untersuchung der Systemleistung in unterschiedlichen Aufgabenbereichen zu. Die Sprachmuster sind verschiedenartigen Einschränkungen unterworfen, die von Sprachschatz bis hin zu situativ bedingten Ausfällen reichen, und die Qualität der Sprachmuster reicht von

☆ This paper is based on a communication presented at the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, Martigny, 5–7 April 1994, and has been recommended by the Scientific Committee of this workshop and the Editorial Board of the journal. This work was sponsored by the Department of the Air Force. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the United States Air Force.

* Corresponding author. Tel.: (617) 981-4494. Fax: (617) 981-0186. E-mail: dar@sst.ll.mit.edu.

Elsevier Science B.V.
SSDI 0167-6393(95)00009-7

nahezu ideal und klar bis hin zu verrauschten Telefonübertragungen. Die Erkennungsgenauigkeit innerhalb abgeschlossener Menge der Sprachmuster in den Datenbanken TIMIT und NTIMIT, die 630 Sprecher umfaßten, betrug 99,5% bzw. 60,7%; bei der Switchboard-Datenbank mit 113 Sprechern betrug sie 82,8%. Bei Verifikationsexperimenten mit der TIMIT-, NTIMIT-, Switchboard- und YOHO-Datenbank ergaben sich globale Schwellenwerte der Fehlerraten von 0,24%, 7,19%, 5,15% bzw. 0,51%.

Résumé

Ce texte présente deux systèmes performants d'identification et de vérification du locuteur fondés sur la modélisation par mélange de gaussiennes, une caractérisation statistique robuste de l'identité d'un locuteur. La méthode d'identification est un classificateur à maximum de vraisemblance; celle de vérification est un test de rapport de vraisemblance appuyé sur une normalisation des locuteurs. Les systèmes ont été évalués sur quatre bases de données publiques de parole: TIMIT, NTIMIT, Switchboard et YOHO. On y trouve une variabilité et des différences de qualité permettant d'évaluer les systèmes selon différents points de vue. Les contraintes y varient de l'élocution par mots isolés à la parole spontanée; la qualité sonore va de la quasi-perfection à l'enregistrement téléphonique. L'identification dans un ensemble fermé de 630 locuteurs, pour TIMIT et NTIMIT, a atteint les taux respectifs de 99.5% et de 60.7%. Pour une population de 113 locuteurs extraits de Switchboard, le taux d'identification a été de 82.8%. Les taux globaux d'erreurs (à seuil égal) de 0.24%, 7.19%, 5.15% et 0.51% ont été obtenus dans les expériences de vérification sur les bases de données TIMIT, NTIMIT, Switchboard et YOHO.

Keywords: Automatic speaker identification and verification; Text-independent; Vocabulary-dependent; Gaussian mixture speaker models; TIMIT; NTIMIT; Switchboard; YOHO

1. Introduction

With the merging of telephony and computers, the growing use of speech as a modality in man–machine communications, and the need to manage speech as a new data-type in multimedia applications, the utility of recognizing a person from his or her voice is increasing. While the area of speech recognition is concerned with extracting the linguistic message underlying a spoken utterance, speaker recognition is concerned with extracting the identity of the person speaking the utterance. Applications of speaker recognition are wide ranging, including facility or computer access control (Naik and Doddington, 1987; Higgins et al., 1991), telephone voice authentication for long-distance calling or banking access (Naik et al., 1989), intelligent answering machines with personalized caller greetings (Schmandt and Arons, 1984), and automatic speaker labeling of recorded meetings for speaker-dependent audio indexing (speech-skimming) (Wilcox et al., 1994; Arons, 1994).

Depending upon the application, the general area of speaker recognition is divided into two specific tasks: identification and verification. In speaker identification, the goal is to determine which one of a group of known voices best matches the input voice sample. This is also referred to as *closed-set* speaker identification. Applications of pure closed-set identification are limited to cases where only enrolled speakers will be encountered, but it is a useful means of examining the separability of speakers' voices or finding similar sounding speakers, which has applications in speaker-adaptive speech recognition. In verification, the goal is to determine from a voice sample if a person is who he or she claims to be. This is sometimes referred to as the *open-set* problem, because this task requires distinguishing a claimed speaker's voice known to the system from a potentially large group of voices unknown to the system (i.e., imposter speakers). Verification is the basis for most speaker recognition applications and the most commercially viable task. The merger of the closed-set identification and open-set verification tasks, called open-set identification, performs like closed-set identification for known speakers but must also be able to classify speakers unknown to the system into a "none of the above" category.

These tasks are further distinguished by the constraints placed on the speech used to train and test the system and the environment in which

the speech is collected (Doddington, 1985). In a text-dependent system, the speech used to train and test the system is constrained to be the same word or phrase. In a text-independent system, the training and testing speech are completely unconstrained. Between text-dependence and text-independence, a vocabulary-dependent system constrains the speech to come from a limited vocabulary, such as the digits, from which test words or phrases (e.g. digit strings) are selected. Furthermore, depending upon the amount of control allowed by the application, the speech may be collected from a noise-free environment using a wideband microphone or from a noisy, narrowband telephone channel.

In this paper a simple but effective statistical speaker representation is presented which attains excellent identification and verification performance for both text-independent and vocabulary-dependent tasks with clean, wideband and telephone speech. The Gaussian mixture speaker model was introduced in (Rose and Reynolds, 1990; Reynolds, 1992) and has demonstrated high text-independent identification accuracy for short test utterances from unconstrained, telephone quality speech. This paper extends the application to speaker verification using background speaker normalization (Higgins et al., 1991) and a likelihood ratio test. A novel technique for selecting background speakers is also presented.

Gaussian mixture model (GMM) based identification and verification systems are evaluated on four publicly available speech databases: TIMIT (Fisher et al., 1986), NTIMIT (Janlowski et al., 1990), Switchboard (Godfrey et al., 1992) and YOHO (Higgins et al., 1991; Campbell, 1992) [1]. Each database possesses different characteristics both in task domain (e.g., text-dependency, number of speakers) and speech quality (e.g., clean wideband, noisy telephone) allowing for experimentation over a wide variety of tasks and conditions. The TIMIT database is used to examine how well text-independent speaker identification can perform under near-ideal conditions with large populations, thus providing an indication of the inherent "crowding" of the feature space. The NTIMIT database is then used to gauge the identification performance loss incurred by transmitting speech over the telephone network for the same large population experiment. The more realistic, unconstrained Switchboard database is used to determine a better measure of large population performance using telephone speech. For speaker verification, the TIMIT, NTIMIT and Switchboard databases are again used to gauge verification performance over the range of near-ideal speech to more realistic, extemporaneous telephone speech. Finally, the YOHO database is used to determine performance on a vocabulary-dependent, office-environment verification task. The effect of different background speaker selections is also examined for all of these databases.

Besides using the databases to address specific research questions, it is hoped that presentation of results on these publicly available databases will encourage competitive evaluations and comparisons by other researchers in the speaker recognition area. There are many competing speaker recognition techniques found throughout the literature, but without evaluation on common databases with defined train/test paradigms it is extremely difficult to assess the merits of an approach. Moreover, few people have the time or resources to implement faithfully a competing scheme to see how it performs on a calibrated database. While not everyone is interested in the same task, the available databases allow evaluation over a wide range of identification and verification scenarios.

The rest of the paper is organized as follows. The next section gives a brief description of the Gaussian mixture speaker model. This is followed in Section 3 by a description of the identification and verification systems. Section 4 then presents descriptions and comparisons of the four databases used in this paper. The identification

[1] Results on the King database with the "great-divide" can be found in (Reynolds, 1994b). TIMIT and NTIMIT are available through the U.S. National Institute of Standards and Technology. Switchboard, YOHO and King are available through the Linguistic Data Consortium.

experimental paradigms and results on these databases are given in Section 5 followed by verification experiments in Section 6. A summary and conclusions are given in Section 7.

2. Gaussian mixture speaker model

The basis for both the identification and verification systems is the GMM used to represent speakers. More specifically, the distribution of feature vectors extracted from a person's speech is modeled by a Gaussian mixture density. For a D-dimensional feature vector denoted as $\boldsymbol{x}$, the mixture density for speaker s is defined as

$$p(\boldsymbol{x}|\lambda_s) = \sum_{i=1}^{M} p_i^s b_i^s(\boldsymbol{x}). \tag{1}$$

The density is a weighted linear combination of M component uni-modal Gaussian densities, $b_i^s(\boldsymbol{x})$, each parameterized by a mean vector, $\boldsymbol{\mu}_i^s$, and covariance matrix, Σ_i^s;

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

The mixture weights, p_i^s, furthermore satisfy the constraint $\sum_{i=1}^{m} p_i^s = 1$. Collectively, the parameters of speaker s's density model are denoted as $\lambda_s = \{p_i^s, \boldsymbol{\mu}_i^s, \Sigma_i^s\}$, $i = 1, \ldots, M$.

While the general model form supports full covariance matrices, in this paper diagonal covariance matrices are used. This choice is based on empirical evidence that diagonal matrices outperform full matrices and the fact that the density modeling of an Mth order full covariance mixture can equally well be achieved using a larger order, diagonal covariance mixture.

Maximum likelihood speaker model parameters are estimated using the iterative Expectation-Maximization (EM) algorithm (Dempster et al., 1977). Generally 10 iterations are sufficient for parameter convergence.

The GMM can be viewed as a hybrid between two effective models for speaker recognition: a uni-modal Gaussian classifier and a vector quantizer codebook. The GMM combines the robustness and smoothness of the parametric Gaussian model with the arbitrary density modeling of the non-parametric VQ model. It can also be viewed as a single-state HMM with a Gaussian mixture observation density or an ergodic Gaussian observation HMM with fixed, equal transition probabilities. Here, the Gaussian components can be considered to be modeling the underlying broad phonetic sounds which characterize a person's voice. A more detailed discussion of how GMMs apply to speaker modeling can be found in (Reynolds, 1992; Reynolds and Rose, 1995).

3. System descriptions

3.1. Speech analysis

Several processing steps occur in the front-end analysis (see Fig. 1). First, the speech is segmented into frames by a 20 ms window progressing at a 10 ms frame rate. A speech activity detector (SAD) is then used to discard silence/noise frames. The SAD is a self-normalizing, energy based detector which tracks the noise floor of the signal and can adapt to changing noise conditions (Reynolds, 1992; Reynolds et al., 1992). For text-independent speaker recognition, it is important to remove silence/noise frames from both the training and testing signal to avoid modeling and detecting the environment rather than the speaker.

Next, mel-scale cepstral feature vectors are extracted from the speech frames (a detailed description of the feature extraction steps can be found in (Reynolds, 1992; Reynolds and Rose, 1995)). For bandlimited telephone speech, cepstral analysis is performed only over the mel-filters in the telephone passband (300–3400 Hz). All

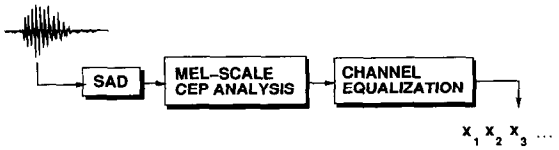

Fig. 1. Front-end speech processing.

cepstral coefficients except $c[0]$ are retained in the processing. This choice of features is based on previous good performance and a recent study (Reynolds, 1994b) comparing several standard speech features for speaker identification.

Last, the feature vectors are channel equalized via blind deconvolution. The deconvolution is implemented by subtracting the average cepstral vector from each input utterance. If training and testing speech are collected from different microphones or channels (e.g., different handsets and/or lines in telephone applications), this is a crucial step for achieving good recognition accuracy (as with the "great-divide" of the King database (Reynolds, 1994b)). However, when there is not much variability between recording microphones or channels, as with the TIMIT/NTIMIT databases, blind channel equalization can reduce accuracy. The channel equalization is used for all databases except the TIMIT and NTIMIT databases.

3.2. Identification system

The identification system is a straight-forward maximum-likelihood classifier. For a reference group of S speakers $\mathcal{S} = \{1,2,\ldots,S\}$ represented by models $\lambda_1, \lambda_2, \ldots, \lambda_S$, the objective is to find the speaker model which has the maximum posterior probability for the input feature vector sequence, $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$. The minimum error Bayes' decision rule for this problem is

$$\hat{s} = \arg \max_{1 \leqslant s \leqslant S} \Pr(\lambda_s | X)$$

$$= \arg \max_{1 \leqslant s \leqslant S} \frac{p(X|\lambda_s)}{p(X)} \Pr(\lambda_s). \tag{3}$$

Assuming equal prior probabilities of speakers, the terms $\Pr(\lambda_s)$ and $p(X)$ are constant for all speakers and can be ignored in the maximum. Using logarithms and the assumed independence between observations, the decision rule becomes

$$\hat{s} = \arg \max_{1 \leqslant s \leqslant S} \sum_{t=1}^{T} \log p(\boldsymbol{x}_t | \lambda_s), \tag{4}$$

in which $p(\boldsymbol{x}_t|\lambda_s)$ is given in Eq. (1). A block diagram of the speaker identification system is shown in Fig. 2(a).

3.3. Verification system

Although requiring only a binary decision, the verification task is more difficult than the identification task in that the alternatives are less defined. The system must decide if the input voice came from the claimed speaker, with a well-defined model, or *not* the claimed speaker, which is ill-defined. Cast in a hypothesis testing framework, for a given input utterance X and a claimed identity the choice is between H_0 and H_1:

H_0: X is from the claimed speaker.

H_1: X is *not* from the claimed speaker.

To perform the optimum likelihood ratio test to decide between H_0 and H_1 then requires some model of the universe of possible non-claimant speakers. The application of this hypothesis testing approach is first described, followed by a discussion of a techniques for selecting speakers for modeling the non-claimant alternative hypothesis.

3.3.1. General approach

The general approach used in the speaker verification system is to apply a likelihood ratio test to an input utterance to determine if the claimed speaker is accepted or rejected. For an utterance $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$ and a claimed speaker identity with corresponding model λ_C, the likelihood ratio is

$$\frac{\Pr(X \text{ is from the claimed speaker})}{\Pr(X \text{ is } \textit{not} \text{ from the claimed speaker})}$$

$$= \frac{\Pr(\lambda_C|X)}{\Pr(\lambda_{\overline{C}}|X)}. \tag{5}$$

Applying Bayes' rule and discarding the constant prior probabilities for claimant and imposter speakers (they are accounted for in the decision threshold), the likelihood ratio in the log domain becomes

$$\Lambda(X) = \log p(X|\lambda_C) - \log p(X|\lambda_{\overline{C}}). \tag{6}$$

The term $p(X|\lambda_C)$ is the likelihood of the utterance given it is from the claimed speaker and $p(X|\lambda_{\overline{C}})$ is the likelihood of the utterance given it

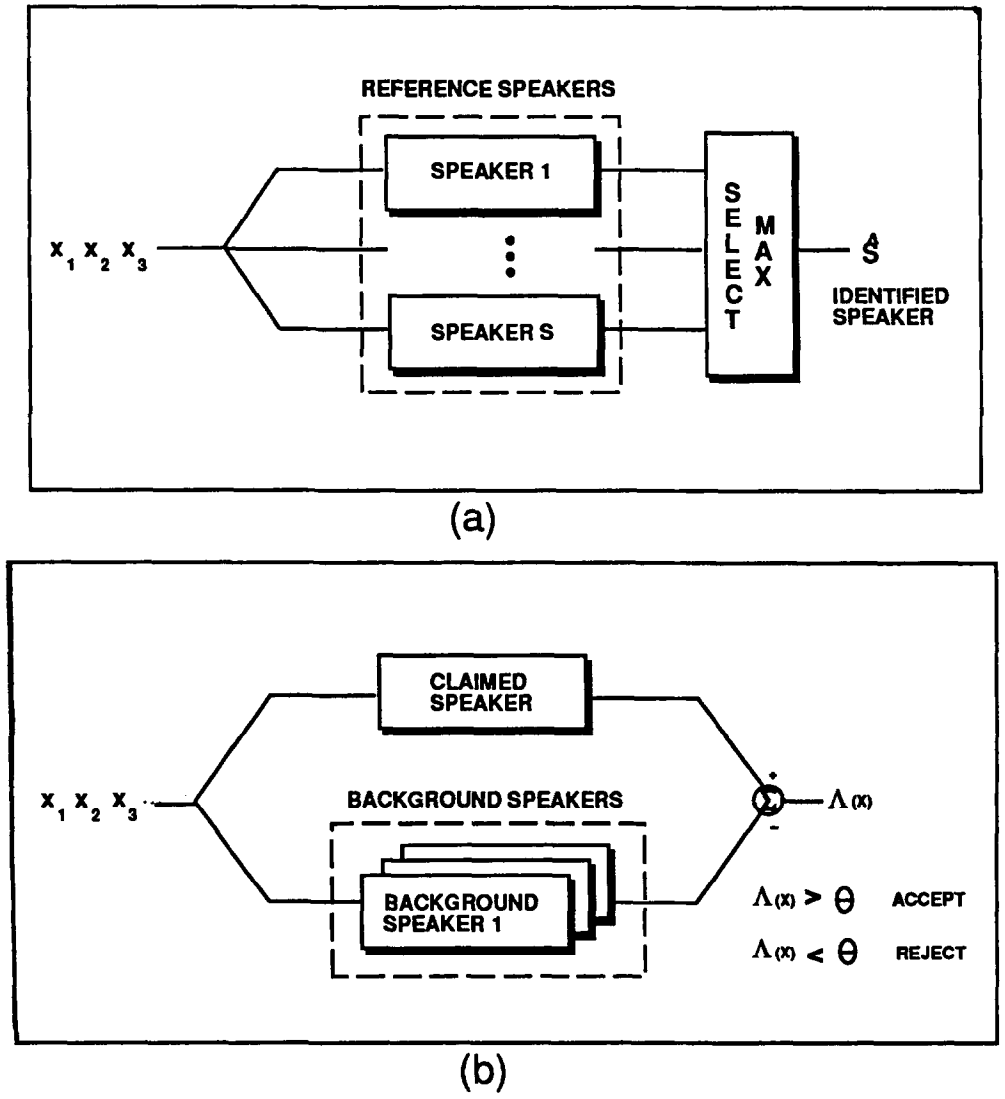

Fig. 2. Speaker recognition systems. (a) Identification system. (b) Verification system.

is not from the claimed speaker. The likelihood ratio is compared to a threshold θ and the claimed speaker is accepted if $\Lambda(X) > \theta$ and rejected if $\Lambda(X) < \theta$. The likelihood ratio essentially measures how much better the claimant's model scores for the test utterance compared to some non-claimant model. The decision threshold is then set to adjust the trade off between rejecting true claimant utterances (false rejection errors) and accepting non-claimant utterances (false acceptance errors).

The terms of the likelihood ratio are computed as follows. The likelihood of the utterance given the claimed speaker's model is directly computed as

$$\log p(X|\lambda_C) = \frac{1}{T}\sum_{t=1}^{T} \log p(\boldsymbol{x}_t|\lambda_C). \tag{7}$$

The $\frac{1}{T}$ scale is used to normalize the likelihood for utterance duration.

The likelihood of the utterance given it is not from the claimed speaker is formed using a collection of background speaker models. With a set of B background speaker models, $\{\lambda_1,\ldots,\lambda_B\}$, the background speakers' log-likelihood is computed as

$$\log p(X|\lambda_{\overline{C}}) = \log\left\{\frac{1}{B}\sum_{b=1}^{B} p(X|\lambda_b)\right\}, \qquad (8)$$

where $p(X|\lambda_b)$ is computed as in Eq. (7). Except for the $\frac{1}{T}$ scale, this is the joint probability density of the utterance coming from one of the background speakers assuming equal-likely speakers. A block diagram of the speaker verification system is shown in Fig. 2(b). The use of background speakers to form various likelihood ratio tests has been used in several different speaker verification systems quite successfully ((Higgins et al., 1991; Rosenberg et al., 1992) for example).

The likelihood normalization provided by the background speakers is important for the verification task because it helps minimize the non-speaker related variations in the test utterance scores, allowing stable decision thresholds to be set. The absolute likelihood score of an utterance from a speaker model is influenced by many utterance-dependent factors including the speaker's vocal characteristics, the linguistic content and the speech quality. These factors make it very difficult to set a decision threshold for absolute likelihood values to be used over different verification tests. The likelihood ratio normalization produces a relative score which is more a function of the utterance speaker and less volatile to non-speaker utterance variations. Note that the identification task does not need the normalization because decisions are made using likelihood scores from a single utterance requiring no inter-utterance likelihood comparisons.

3.3.2. Background speaker selection

Two issues that arise with the use of background speakers are the selection of the speakers and the number of speakers to use. Intuitively, the background speakers should be selected to represent the population of expected imposters, which is in general application specific. In some scenarios, it may be assumed that imposters will attempt to gain access only from similar sounding or at least same-sex speakers (dedicated imposters). In a telephone based application accessible by a larger cross-section of potential imposters, on the other hand, the imposters may sound very dissimilar to the users they attack (casual imposters); for example a male imposter claiming to be a female user. Previous systems have relied on selecting background speakers whose models are "closest" or most competitive for each enrolled speaker (termed the ratio-set or cohorts). This may be appropriate for the dedicated imposter scenario, but, as seen in the experiments and discussed in (Higgins et al., 1991), this leaves the system vulnerable to imposters which have very dissimilar voice characteristics. This occurs because the dissimilar voice is not modeled well by either the numerator or denominator of the likelihood ratio. The test is based on tails of distributions of the speaker models giving rise to unreliable values. Although it is possible to employ methods of rejecting very dissimilar voices based on thresholding the probability score from the claimed speaker's model (Higgins et al., 1991), the approach of judicious background speaker selection is pursued here. The experiments presented in Section 6 examine both the same-sex and mixed-sex imposter situations.

Ideally the number of background speakers should be as large as possible to better model the imposter population, but practical considerations of computation and storage dictate a small set of background speakers. In the verification experiments, the number of background speakers is set to ten. The limited size was motivated by real-time computation considerations and the desire to set a constant experimental test. For a verification experiment on a given database, each speaker is used as a claimant with the remaining speakers (excluding the claimant's background speakers) acting as imposters and rotating through all speakers. Using large background speaker sets decreases the number of imposter tests. This paradigm allows the maximum use of the speakers for a large number of tests but makes it

difficult to compare systems using different size background speaker sets. An alternative approach used in the Switchboard verification experiments is to partition the database speakers into separate claimant and imposter groups with background speakers selected only from the claimant group.

For the dedicated imposter case, the selection of background speakers for each claimant speaker is done as follows. Using the training data, GMMs of all speakers in the database are created and pair-wise distances between the speaker models are computed. For speakers i and j with models (λ_i, λ_j) and training utterances (X_i, X_j), the distance is defined as

$$d(\lambda_i,\lambda_j) = \log\frac{p(X_i|\lambda_i)}{p(X_i|\lambda_j)} + \log\frac{p(X_j|\lambda_j)}{p(X_j|\lambda_i)}. \quad (9)$$

The ratio $p(X_i|\lambda_i)/p(X_i|\lambda_j)$ measures how well speaker j's model scores with speaker i's speech relative to how well speaker i's model scores with his/her own speech. The more similar the models, the smaller the ratio becomes. The distance measure is then a symmetric combination of ratios comparing models λ_i and λ_j.

Each speaker's N closest speakers [2] ($N > B$, where B is the size of the final background speaker set) are then selected as his/her "close cohort" denoted as $\mathscr{C}(i)$ for speaker i. From $\mathscr{C}(i)$, the final B background speaker set, denoted $\mathscr{B}(i)$, is selected by finding those B which are maximally spread from each other. More specifically, the procedure is as follows:

(0) Start by moving the closest speaker from $\mathscr{C}(i)$ to $\mathscr{B}(i)$. $N = N - 1$, $B' = 1$ (B' is the current number of speakers in $\mathscr{B}(i)$).

(1) Move speaker c from $\mathscr{C}(i)$ to $\mathscr{B}(i)$, where c is found by

$$c = \arg\max_{c\in\mathscr{C}(i)} \left\{ \frac{1}{B'} \sum_{b\in\mathscr{B}(i)} \frac{d(\lambda_b,\lambda_c)}{d(\lambda_i,\lambda_c)} \right\}.$$

$N = N - 1$, $B' = B' + 1$

(2) Repeat step (1) until $B' = B$.

The maximal spread constraint is used to eliminate "duplicate" background speakers which are similar to each other and so obtain the best coverage for the limited number of selected speakers. The background speakers selected as above are referred to as the maximally-spread close (msc) set.

When the imposter population contains dissimilar speakers from the users (such as opposite sex imposters), the background selection should also include far (dissimilar) as well as close speakers. The background speaker set in this case is equally divided between the close and far speakers. The far speaker selection is accomplished by using the pair-wise distance measure as before, but now selecting the maximally spread $B/2$ speakers from the N farthest speakers. In particular, each speaker's N farthest speakers are selected as his/her "far cohort" denoted as $\mathscr{F}(i)$ for speaker i. The maximally spread $B/2$ background speakers are selected via the following procedure:

(0) Start by moving the farthest speaker from $\mathscr{F}(i)$ to $\mathscr{B}(i)$. $N = N - 1$, $B' = 1$.

(1) Move speaker f from $\mathscr{F}(i)$ to $\mathscr{B}(i)$, where f is found by

$$f = \arg\max_{f\in\mathscr{F}(i)} \left\{ \frac{1}{B'} \sum_{b\in\mathscr{B}(i)} d(\lambda_b,\lambda_f) * d(\lambda_i,\lambda_f) \right\}.$$

$N = N - 1$, $B' = B' + 1$

(2) Repeat step (1) until $B' = B/2$.

These dissimilar background speakers are referred to as the maximally-spread far (msf) set.

4. Database description

Some relevant characteristics of the databases used in the experiments are shown in Table 1. More details about the databases' characteristics and collection set-ups can be found in their references.

The TIMIT database allows examination of speaker identification performance under almost ideal conditions. With the 8 kHz bandwidth, and

[2] $N = 20$ and $B = 10$ for the verification experiments.

Table 1
Characteristics of databases used (PSTN = Public Switched Telephone Network)

Database	# speakers	# utterances/ speaker	Channel	Acoustic environment	Handset	Intersession interval
TIMIT (Fisher et al., 1986)	630	10 read sentences	clean	sound booth	wideband microphone	none
NTIMIT (Janlowski et al., 1990)	630	10 read sentences	PSTN local and long distance	sound booth	fixed carbon button	none
Switchboard (Godfrey et al., 1992)	500	1–25 conversation	PSTN long distance	home and office	variable	variable days–weeks
YOHO (Higgins et al., 1991)	138	4/train 10/test combination lock	clean	office	telephone hi-quality microphone	variable days–months

because of the lack of intersession variability, acoustic noise and microphone variability and distortion, recognition errors should be a function of overlapping speaker distributions. Furthermore, each utterance is a read sentence of approximately 3 seconds duration. The sentences are designed to have rich phonetic variability. Note that this is a factor which favorably biases TIMIT performance compared to 3 second length utterances extracted at random from extemporaneous speech.

The NTIMIT database is the same speech from the TIMIT database recorded over local and long-distance telephone loops. Each TIMIT sentence was played through an "artificial mouth" coupled to a carbon-button telephone handset via a telephone test frame designed to approximate the acoustic coupling between the human mouth and the telephone handset. The speech was transmitted through a local or long-distance central office and looped back for recording. This provides the identical TIMIT speech, but degraded through carbon-button transduction and actual telephone line conditions. Performance differences between identical experiments on TIMIT and NTIMIT should arise mainly from the effects of the microphone and telephone transmission degradations.

The Switchboard database provides one of the best telephone speech, speaker recognition databases available. Large amounts of spontaneous telephone speech from hundreds of speakers collected under home/office acoustic conditions with varying telephone handsets make recognition results from Switchboard more realistic for telephone based applications. The channel conditions tend to be clean so that channel noise is not a major issue. However, background noise from radios or televisions can be found in some recordings. Each side of a two-way conversation was recorded separately to allow isolation of single speaker speech. However, due to limits of the telephone network echo cancelling performance, even single conversation halves may have low-level opposite channel echo present. In this work, speaker turns from the transcripts and differential-energy echo suppression were used to isolate single speaker speech for training and testing.

The YOHO database was designed to support text-dependent speaker verification research such as is used in secure access technology. It has a well defined train/test scenario in which each speaker has four enrollment sessions where he/she is prompted to read a series of 24 combination-lock phrases. Each phrase is a sequence of three two-digit numbers (e.g., "35–72–41", pronounced "thirty-five seventy-two forty-one"). There are 10 verification trials per speaker consisting of four phrases per trial. The vocabulary consists of 56 two-digit numbers ranging from 21 to 97 (see (Higgins et al., 1991) for the selection rules). The speech was collected in an office environment using a telephone handset connected to a workstation. Thus, the data has a telephone bandwidth of 3.8 kHz, but no telephone transmission degradations. The YOHO database is different from the above text-independent, telephone speech databases and allows the demonstration of how the GMM verification system, although designed for text-independent operation, can also perform very well under the vocabulary-dependent constraints of this application.

5. Identification experiments

Closed-set identification experiments were conducted on the TIMIT, NTIMIT and Switchboard databases. The goal of the experiments was to examine the performance of the identification system as a function of population size for both clean, wideband speech and telephone speech. The TIMIT performance provides an indication of how crowded the feature space is under near-ideal conditions. The NTIMIT results indicate the performance loss from using noisy, telephone speech. Results on the more realistic Switchboard database provide a better measure of expected extemporaneous, telephone speech performance and the effect of handset variability.

5.1. TIMIT / NTIMIT results

For the identification experiments on the TIMIT and NTIMIT databases, all 630 speakers

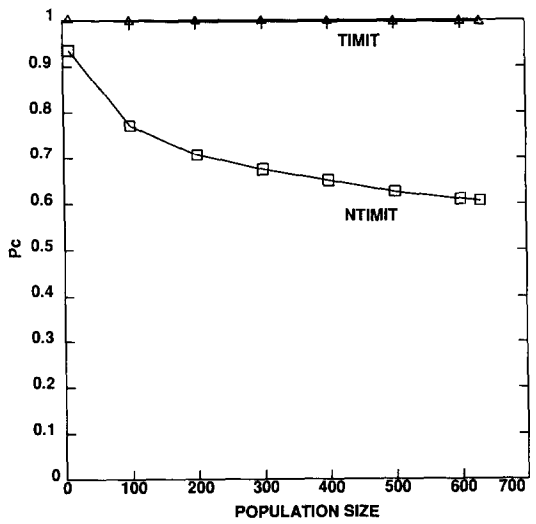

Fig. 3. Speaker identification accuracy as a function of population size on TIMIT and NTIMIT databases. Thirty-two component GMMs trained with 24 seconds of speech and tested with 3 second utterances.

(438 males, 192 females) were used. Speaker models with 32 Gaussians were trained using eight utterances of approximately 24 seconds total duration. The remaining two utterances of approximately 3 seconds duration each were individually used as tests (a total of 1260 tests) [3].

Identification accuracy for a population size S was computed by performing repeated speaker identification experiments on 50 sets of S speakers randomly selected from the total pool of 630 available speakers and averaging the results. This helps average out the bias of a particular population composition. Population sizes of (10, 100, 200, 300, 400, 500, 600, 630) were used. The results are shown in Fig. 3.

Under the near-ideal TIMIT conditions, performance is barely affected by increasing population sizes. This indicates that the limiting factor in speaker identification performance is not a crowding of the feature space. However, with telephone line degradations, the NTIMIT accuracy steadily decreases as population size increases. The largest drop in accuracy occurs as the population size increases to 100. Above 200 speakers the accuracy decrease becomes almost linear. With the full 630 speaker populations, there is a gap of 39 percentage points between TIMIT and NTIMIT accuracy (TIMIT Pc = 99.5%, NTIMIT Pc = 60.7%). The correct TIMIT speakers had an average rank of 1.01, while the correct NTIMIT speakers had an average rank of 8.29 [4].

For the complete 630 population TIMIT database, there were no cross-sex errors, with male and female accuracies of 99.8% and 99.0%, respectively. On the complete 630 population NTIMIT database, there were four cross-sex errors. Male speaker accuracy was 62.5% versus 56.5% for female speakers. It is not clear why the female speakers have a lower accuracy than the male speakers.

Examining the NTIMIT database, the main degradation appears to be noise and bandlimiting. The TIMIT database has an average signal-to-noise ratio (SNR) [5] of 53 dB, while the NTIMIT database has an average SNR of 36 dB. Examination of sweep tones from each telephone line used in the NTIMIT database shows little spectral shape variability. This is not surprising since the telephone handset is the source of most spectral shaping and a single handset was used for all recordings. Detailed studies systematically imposing various degradations on TIMIT speech (e.g., bandlimiting, noise addition) to explain the performance gap between the TIMIT and NTIMIT databases can be found in (Reynolds, 1994a; Reynolds et al., 1995).

Using a different training and testing paradigm, recently published results on the complete 630 speaker TIMIT database in (Floch et al., 1994) also show a very high accuracy of 95.6% using a text-independent technique which scores only se-

[3] The eight training utterances were the 2 sa sentences, 3 si sentences and first 3 sx sentences. The two testing utterances were the remaining 2 sx sentences.

[4] A speaker's rank for a test utterance is the position of his model's score within the sorted list of speaker model scores; a rank of 1 being the best scoring speaker.

[5] SNR is computed using the ratio of signal-peak energy to noise-floor energy over the entire utterance.

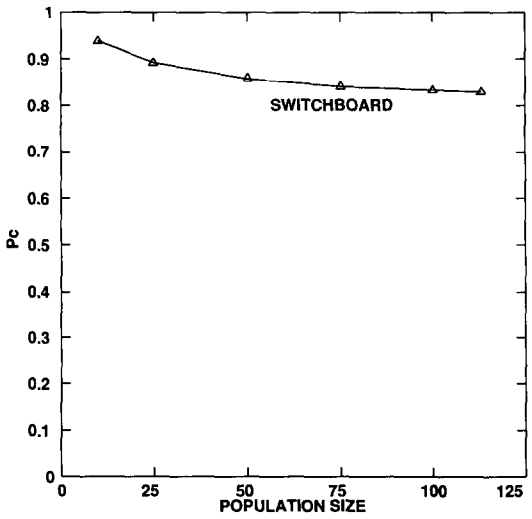

Fig. 4. Speaker identification accuracy as a function of population size on Switchboard database. Sixty-four component GMMs trained with 6 minutes of speech and tested with 1 minute utterances.

lected phonetic clusters. To the author's knowledge, there have been no published speaker identification experiments conducted on the complete NTIMIT database.

5.2. Switchboard results

For the Switchboard database, 113 speakers (50 males, 63 females) were used with 64 component GMMs trained using 6 minutes of speech extracted equally from 2 conversations. Testing was done on a total of 472 utterances of 1 minute duration. There were 2–12 test utterances per speaker with an average of four utterances. Identification accuracy was computed as above, except 100 sets per population size were used for population sizes of (10, 25, 50, 75, 100, 113). Results are shown in Fig. 4.

Although not directly comparable, the Switchboard results exhibit the same decreasing trend as the NTIMIT results, but not as rapidly. Due to the increased training and testing data and the higher SNRs (typically 40 dB or higher), the Switchboard results are better than the NTIMIT results. For the 113 population, the overall accuracy was 82.8% with an average rank of 2.29. There were two cross-sex errors and the male speakers had an accuracy of 81.0% compared to the female speaker accuracy of 84.3%.

The effect of handset variability on the results was examined using the telephone numbers associated with the training and testing utterances. For each conversation in the Switchboard database, a coded version of the callers' telephone numbers are given. Conversations originating from identical telephone numbers can generally be assumed to be over the same telephone handset. Conversely, it may be assumed that there is a correlation between conversations originating from different telephone numbers and the caller using different handsets [6]. There are, of course, other factors, such as different transmission paths and acoustic environments, which also change with different telephone numbers. The aim here is to examine the performance when training and testing utterances originate from the same and different telephone numbers under the assumption that the telephone number implies a handset.

Since the speaker models were trained from two conversations, there are at most two training telephone numbers (handsets) per speaker. Of the 113 speakers, 95 had both of their training utterances from the same telephone number. The first row in Table 2 shows the number of test utterances with and without train/test telephone number matches. A train/test match occurs if a speaker's testing utterance had the same telephone number as either of the training utterances. There is a clear dominance in this test of matched telephone numbers. The second row of the table shows the number of misclassifications for the two groups. Here it is seen that most of the errors are from the mismatched conditions, with 45% of the total number of errors coming from the mismatched group comprising only 16% of the total number of tests. The error rate in the mismatched group is almost five times that in the

[6] Neither assumption is strictly true, since callers can use different telephone units with the same telephone number and similar telephone units can be used at different telephone numbers.

Table 2
Number and accuracy of testing utterances with and without matching telephone numbers to training utterances for Switchboard identification experiment

	No matching telephone numbers	At least one matching telephone number
Number of test utterances	74	398
Number of errors (percent error)	35 (47.3%)	43 (10.8%)

matched group indicating the sensitivity to acoustic mismatches between training and testing conditions. Given that so many mismatch errors occurred even with blind deconvolution channel equalization, further indicates that the degradations are more complex than a first-order linear filter effect.

Other published speaker identification results on the Switchboard database typically use a smaller 24 speaker set (12 male, 12 female) with a total of 97 test utterances (1–6 utterances per speaker). On this task, using 10 and 60 second long test utterances, the GMM system has an accuracy of 94% @ 10 s and 95% @ 60 s compared to 96% @ 60 s for ITT's nearest neighbor classifier (Higgins et al., 1993), 90% @ 10 s and 95% @ 60 s for BBN's Gaussian classifier (Gish and Schmidt, 1994), and 89% @ 10 s and 88% @ 60 s for Dragon System's continuous speech recognition classifier (Gillick et al., 1993) [7]. Using robust scoring techniques, the accuracy can be increased to almost 100% for both utterance lengths (Gish and Schmidt, 1994; Bahler et al., 1994). As above, there was significant overlap between training and testing telephone handsets which favorably biases performance.

6. Verification experiments

Verification experiments were conducted on the TIMIT, NTIMIT, Switchboard and YOHO databases. The TIMIT, NTIMIT and Switchboard databases are again used to gauge verification performance over the range of near-ideal speech to more realistic, extemporaneous telephone speech. The YOHO database is used to demonstrate performance for a vocabulary-dependent, office-environment, secure access application.

As discussed in Section 3.3.2, the composition of the imposter speakers can greatly affect performance. Experiments using same-sex imposters and mixed-sex imposters are presented in conjunction with the two different background speaker selection procedures (see Section 3.3.2). There are two same-sex experiments and one mixed-sex experiment: one using male speakers only (M), one using female speakers only (F) and one using male and female speakers (M + F) together. Two background speaker sets of size ten, 10 maximally-spread close (10 msc) and 5 maximally-spread close plus 5 maximally-spread far (5 msc, 5 msf), were selected from the complete set of speakers for each database. Since the msf speakers are selected from the complete database, they generally represent opposite sex speakers. In all experiments, the background speaker utterances were excluded from the imposter tests.

Results are reported as the equal-error rate (eer) computed using a global threshold. This is found by placing all the true test scores and imposter test scores in one sorted list and locating the point on the list at which the false-acceptance (FA) rate (percent of imposter tests above the point) equals the false-rejection (FR) rate (the percent true tests below the point); the eer is the FA rate at this point. The global threshold eer measures the overall (speaker independent) system performance using the largest number of true and imposter tests available. Results using speaker dependent thresholds (i.e., treating each speaker's true and imposter utterances scores separately) will generally be higher than global threshold results, but may have lower statistical significance due to using the smaller number of

[7] The testing paradigm is the same for these systems, but the training paradigm is not.

tests available per speaker. The issues of independence between tests and statistical significance of results on each database are difficult questions which have not been clearly addressed in the literature.

6.1. TIMIT / NTIMIT results

For the verification experiments on TIMIT and NTIMIT, the 168 speakers (112 males, 56 females) from the "test" portion of the databases were used. As in the identification experiment, speaker models with 32 Gaussians were trained using eight utterances of approximately 24 seconds total duration. The remaining two utterances of approximately 3 seconds duration each were individually used as tests. Experiments were performed using each speaker as a claimant with the remaining speakers (excluding the claimant's background speakers) acting as imposters and rotating through all speakers. The number of claimant tests and imposter tests for the M, F and M + F tests are given in Table 3.

Results for the three experimental conditions (M, F, M + F) and two background speaker selections are given in Table 4. As with the speaker identification results, almost perfect performance is obtained on the TIMIT database, with the NTIMIT performance doing significantly worse. The NTIMIT best M + F eer is about 30 times worse than the TIMIT eer. Comparing the M + F experiments with and without the far background speakers, it is clear that inclusion of the dissimilar speakers improved performance by better modeling the imposter population. As expected, the dissimilar speakers for the male speakers were mainly female speakers and vice versa. However, since there is a predominance of male speakers in the M + F test, the improvement was not as great as may have occurred with a more balanced test. As in the identification tests, the male speakers are performing better than the female speakers.

6.2. Switchboard results

The verification paradigm on the Switchboard database is different from that used on the TIMIT and NTIMIT databases. Here, 24 claimant speakers (12 males, 12 females) were each represented by 64 component GMMs trained using 3 minutes of speech extracted equally from four conversations. There were a total of 97 claimant utterances of 16 seconds average duration selected from conversations. Claimants had between one and six true tests with an average of four. A separate set of 428 utterances of 16 seconds average duration from 210 speakers (99 males and 111 females) was used for the imposter tests. The utterances were designated using speaker-turns from the transcripts so as to isolate single-speaker speech. The number of claimant and imposter

Table 3
Number of claimant and imposter trials for TIMIT and NTIMIT databases. Background speaker set size of 10

Experiment	# speakers	# true tests per speaker	# imposter tests per speaker	Total # true tests	Total # imposter tests
M	112	2	202	224	22624
F [a]	56	2	88	110	4945
M + F	168	2	313	334	52538

[a] Two speakers had only one test utterance.

Table 4
Equal error rate (%) for same-sex (M, F) and mixed-sex (M + F) experiments on TIMIT and NTIMIT databases (msc = maximally-spread close background speakers, msf = maximally-spread far background speakers)

Database	M (10 msc)	M (5 msc, 5 msf)	F (10 msc)	F (5 msc, 5 msf)	M + F (10 msc)	M + F (5 msc, 5 msf)
TIMIT	0.14	0.32	0.28	0.71	0.50	0.24
NTIMIT	8.15	8.48	8.79	10.44	8.68	7.19

tests for the M, F and M + F experiments are given in Table 5.

Two background speaker sets were used from this relatively small claimant population: a same-sex set (ss), in which each speaker used all other claimant speakers of the same sex as background speakers, and a selection consisting of five maximally-spread close and five maximally-spread far background speakers (essentially a mixed-sex set). The results for these experiments are shown in Table 6.

It is first surprising to see that the same-sex background set (11 ss) did worse that the mixed-sex background set (5 msc, 5 msf) on the M and F experiments. Since same-sex imposters were used in these tests, it was expected that using same-sex background speakers would perform better than a background set split between males and females. However, closer examination of the utterances in error found that they generally were extracted from a mixed-sex conversation and that the echo from the opposite side was contaminating the utterance. Thus, for example, some ostensibly male only imposter utterances actually contained female speech. As with the TIMIT and NTIMIT experiments, a decrease in eer is obtained in the M + F experiment using the mixed-sex (close and far) background speaker set.

Examination of the claimant training and testing utterance telephone numbers also found only 16% of the claimant tests were from telephone numbers unseen in the training data which favorably biases the FR rate. In the mismatched cases, some speakers had very high FR errors.

6.3. YOHO results

For the YOHO experiments, each speaker was modeled by a 64 component GMM trained using the four enrollment sessions (average of 6 minutes). Each speaker has 10 verification sessions consisting of four combination-lock phrases (average of 15 seconds). Experiments were performed using each speaker as a claimant with the remaining speakers (excluding the claimant's background speakers) acting as imposters and rotating

Table 5
Number of claimant and imposter trials for Switchboard database. Separate claimant and imposter populations used

Experiment	# speakers	Average # true tests per speaker	# imposter tests per speaker	Total # true tests	Total # imposter tests
M	12	4	210	47	2520
F	12	4	218	50	2616
M + F	24	4	428	97	10272

Table 6
Equal error rate (%) for same-sex (M, F) and mixed-sex (M + F) experiments on Switchboard databases (ss = same sex background speakers, msc = maximally-spread close background speakers, msf = maximally-spread far background speakers)

Database	M (11 ss)	M (5 msc, 5 msf)	F (11 ss)	F (5 msc, 5 msf)	M + F (11 ss)	M + F (5 msc, 5 msf)
Switchboard	5.83	4.25	11.39	7.99	8.25	5.15

Table 7
Number of claimant and imposter trials for YOHO database. Background speaker set size of 10

Experiment	# speakers	# true tests per speaker	# imposter tests per speaker	Total # true tests	Total # imposter tests
M	106	10	950	1060	100700
F [a]	32	10	210	318	6720
M + F	138	10	1268	1378	175105

[a] Two speakers had only nine test sessions.

Table 8
Equal error rate (%) and false rejection rates at false acceptance rates of 0.1% and 0.01% for same-sex (M, F) and mixed-sex (M + F) experiments on YOHO database (msc = maximally-spread close background speakers, msf = maximally-spread far background speakers)

Database	M (10 msc)	M (5 msc, 5 msf)	F (10 msc)	F (5 msc, 5 msf)	M + F (10 msc)	M + F (5 msc, 5 msf)
YOHO (eer)	0.20	0.28	1.88	1.57	0.58	0.51
(FR @ FA = 0.1%)	0.38	0.38	1.89	1.89	0.87	0.65
(FR @ FA = 0.01%)	0.94	2.36	2.51	3.77	2.40	2.40

through all speakers. As with the TIMIT and NTIMIT databases, there is a gender imbalance with 106 male speaker and only 32 female speakers. The number of claimant tests and imposter tests for the M, F and M + F tests are given in Table 7.

Results for the three experimental conditions using the two background speaker sets are given in Table 8. In addition to the eer, the table also gives the false rejection rate at a false acceptance rate of 0.1% and 0.01%. These latter numbers are given to measure performance at tight operating specification for an access control application. It is clear that very low error rates are achievable for this task, owing to the good quality and vocabulary constraints of the speech. The vocabulary constraints mean that a speaker's GMM need only model a constrained acoustic space thus allowing an inherently text-independent model to effectively use the text-dependent training and testing data. The high performance is also found for closed set identification using the same data: accuracy for males of 99.7%, for females of 97.8% and for males and females of 99.3%. As with the other databases, performance is lower for the female speakers. The close and far background selection slightly boosted performance for the M + F experiment, which again is dominated by male speakers.

In (Campbell, 1995), Campbell presents verification and identification results on the YOHO database from several different systems. Comparing to the GMM system's eer of 0.5%, ITT's continuous speech recognition classifier has an eer of 1.7% (Higgins et al., 1991), ITT's nearest neighbor classifier has an eer of 0.5% (Higgins et al., 1992), and Rutgers University's neural tree network has an eer of 0.7% (Liou and Mammone, 1995). These results can only be loosely compared, however, since different training/testing paradigms and background speaker sets were used (e.g., ITT's CSR system used five background speakers. The Rutgers' system had overlap between background and imposter speakers).

7. Conclusion

This paper has presented and evaluated identification and verification systems for text-independent speaker recognition based on Gaussian mixture speaker models. The GMM provides a simple yet effective speaker representations which is computationally inexpensive and provides high recognition accuracy. Using this probabilistic speaker model, the recognition systems were defined as implementations of maximum likelihood classification and hypothesis testing rules.

Experimental evaluation of the systems' performance was conducted on four publicly available speech databases: TIMIT, NTIMIT, Switchboard and YOHO. Each database offers different levels of speech quality and control. The TIMIT database provides near-ideal speech with clean, wideband quality recordings, no inter-session variabilities, and phonetically rich, read speech. Under these ideal conditions, it was determined that speaker crowding of the feature space was not an issue for population sizes up to 630. A closed set identification accuracy of 99.5% was achieved for the complete 630 speaker population. The NTIMIT database adds real telephone line degradations to the TIMIT data and it was found that these degradations caused large per-

formance losses. The NTIMIT accuracy dropped to 60.7% for the same 630 population identification task. For verification, the TIMIT eer was 0.24% compared to 7.19% on NTIMIT.

The Switchboard database provides the most realistic mix of real-world variabilities which can affect speaker recognition performance. The performance trends on Switchboard appeared similar to those found with NTIMIT, producing an 82.8% identification accuracy for a 113 speaker population and an eer of 5.15% for a 24 speaker verification experiment. The factors degrading the NTIMIT and Switchboard performances, however, are different. High noise levels seem to be the main degradation in NTIMIT, whereas handset variability and cross-channel echo are the two major degradations in Switchboard. For the identification experiments, it was found that the error rate for utterances from telephone numbers unseen in the training utterances was almost five times that of utterances from telephone numbers found in the training utterances.

Finally, results on the YOHO database show that very low error rates are possible for a secure access verification application even using a text-independent verification system. An overall eer of 0.51% and a false rejection rate of 0.65% at a 0.1% false acceptance rate was obtained. The constrained vocabulary along with the good quality speech allowed the model to focus in on the sounds which characterize a person's voice without extraneous channel variabilities.

The verification experiments also demonstrated the need to select background speakers to cover the population of expected imposter speakers. A procedure for selecting maximally-spread close and far background speakers was described. For the mixed-sex experiments, a background speaker set equally split between close and far speakers gave better performance than using only close speakers. It was also observed from the same-sex experiments, that the verification system performed better for the male speakers than for the female speakers (also observed in closed-set TIMIT/NTIMIT and YOHO results).

The experimental results presented in this paper indicate that the major limiting factor in performance is transmission degradations, including noise and microphone variability. Any future gains under these conditions will most likely come from application of robustness techniques to both the classifier and the front-end analysis.

Acknowledgements

The author wishes to thank Jerry O'Leary, Marc Zissman, Doug Paul, Cliff Weinstein, Richard Lippmann and others in the SST Group at Lincoln for many helpful discussions and assistance throughout this work.

References

B.M. Arons (1994), Interactively skimming recorded speech, PhD thesis, Massachusetts Institute of Technology, February.

L.G. Bahler, J.E. Porter and A.L. Higgins (1994), "Improved voice identification using a nearest-neighbor distance measure", *Proc. Internat. Conf. Acoust. Speech Signal Process., April 1994*, pp. I-321–I-323.

J.P. Campbell (1992), Features and measures for speaker recognition, PhD thesis, Oklahoma State University, December.

J.P. Campbell (1995), "Testing with the YOHO CD-ROM voice verification corpus", *Proc. Internat. Conf. Acoust. Speech Signal Process., May 1995*, pp. 341–344.

A. Dempster, N. Laird and D. Rubin (1977), "Maximum likelihood from incomplete data via the EM algorithm", *J. Roy. Statist. Soc.*, Vol. 39, pp. 1–38.

G. Doddington (1985), "Speaker recognition-identifying people by their voices", *Proc. IEEE*, Vol. 73, pp. 1651–1664.

W.M. Fisher, G.R. Doddington and K.M. Goudie-Marshall (1986), "The DARPA speech recognition research database: Specifications and status", *Proc. DARPA Workshop on Speech Recognition, February 1986*, pp. 93–99.

J.L. Floch, C. Montacie and M.J. Caraty (1994), "Investigations on speaker characterization from Orphee system technics", *Proc. Internat. Conf. Acoust. Speech Signal Process., April 1994*, pp. I-149–I-152.

L. Gillick et al. (1993), "Application of large vocabulary continuous speech recognition to topic and speaker identification using telephone speech", *Proc. Internat. Conf. Acoust. Speech Signal Process., April 1993*, pp. II-471–II-474.

H. Gish and M. Schmidt (1994), "Text-independent speaker identification", *IEEE Signal Processing Mag.*, Vol. 11, pp. 18–32.

J.J. Godfrey, E.C. Holliman and J. MacDaniel (1992), "Switchboard: Telephone speech corpus for research and

development", *Proc. Internat. Conf. Acoust. Speech Signal Process., March 1992,* pp. I-517–I-520.

A. Higgins, L. Bahler and J. Porter (1991), "Speaker verification using randomized phrase prompting", *Digital Signal Processing,* Vol. 1, pp. 89–106.

A. Higgins, L. Bahler, G. Vensko, J. Porter and D. Vermilyea (1992), YOHO speaker authentication final report, Technical Report, ITT Defense Communications Division.

A. Higgins, L. Bahler and J. Porter (1993), "Voice identification using nearest-neighbor distance measure", *Proc. Internat. Conf. Acoust. Speech Signal Process. 1993,* pp. II-375–II-378.

C. Jankowski, A. Kalyanswamy, S. Basson and J. Spitz (1990), "NTIMIT: A phonetically balanced, continuous speech telephone bandwidth speech database", *Proc. Internat. Conf. Acoust. Speech Signal Process., April 1990,* pp. 109–112.

H-S. Liou and R. Mammone (1995), "A subword neural tree network approach to text-dependent speaker verification", *Proc. Internat. Conf. Acoust. Speech Signal Process., May 1995,* pp. 357–360.

J. Naik and G. Doddington (1987), "Evaluation of a high performance speaker verification system for access control", *Proc. Internat. Conf. Acoust. Speech Signal Process., April 1987,* pp. 2392–2395.

J. Naik, L. Netsch and G. Doddington (1989), "Speaker verification over long distance telephone lines", *Proc. Internat. Conf. Acoust. Speech Signal Process. 1989,* pp. 524–527.

D.A. Reynolds (1992), A Gaussian mixture modeling approach to text-independent speaker identification, PhD thesis, Georgia Institute of Technology, September.

D.A. Reynolds, R.C. Rose and M.J.T. Smith (1992), "PC-based TMS320C30 implementation of the Gaussian mixture model text-independent speaker recognition system", *Proc. Internat. Conf. Signal Processing Applications and Technology, November 1992,* pp. 967–973.

D.A. Reynolds (1994a), "Effects of population size and telephone degradations on speaker identification performance", *Proc. SPIE Conf. on Automatic Systems for the Identification and Inspection of Humans, July 1994.*

D.A. Reynolds (1994b), "Experimental evaluation of features for robust speaker identification", *IEEE Trans. Speech and Audio Processing,* Vol. 2, pp. 639–643.

D.A. Reynolds et al. (1995), "The effects of telephone transmission degradations on speaker recognition performance", *Proc. Internat. Conf. Acoust. Speech Signal Process., May 1995,* pp. 329–332.

D.A. Reynolds and R.C. Rose (1995), "Robust text-independent speaker identification using Gaussian mixture speaker models", *IEEE Trans. Speech and Audio Processing,* Vol. 3, pp. 72–83.

R.C. Rose and D.A. Reynolds (1990), "Text-independent speaker identification using automatic acoustic segmentation", *Proc. Internat. Conf. Acoust. Speech Signal Process.,* pp. 293–296.

A.E. Rosenberg, J. DeLong, C.H. Lee, B.H. Juang and F.K. Soong (1992), "The use of cohort normalized scores for speaker verification", *Internat. Conf. Speech and Language Processing,* November, pp. 599–602.

C. Schmandt and B. Arons (1984), "A conversational telephone messaging system", *IEEE Trans. Consumer Electronics,* Vol. 30, pp. xxi–xxiv.

L. Wilcox, F. Chen, D. Kimber and V. Balasubramanian (1994), "Segmentation of speech using speaker identification", *Proc. Internat. Conf. Acoust. Speech Signal Process.,* pp. I-161–I-164.