Discriminative In-Set/Out-of-Set Speaker Recognition

Pongtep Angkititrakul, *Member, IEEE*, and John H. L. Hansen, *Fellow, IEEE*

***Abstract*—In this paper, the problem of identifying in-set versus out-of-set speakers for limited training/test data durations is addressed. The recognition objective is to form a decision regarding an input speaker as being a legitimate member of a set of enrolled speakers or outside speakers. The general goal is to perform rapid speaker model construction from limited enrollment and test size resources for in-set testing for input audio streams. In-set detection can help ensure security and proper access to private information, as well as detecting and tracking input speakers. Areas of applications of these concepts include rapid speaker tagging and tracking for information retrieval, communication networks, personal device assistants, and location access. We propose an integrated system with emphasis on short-enrollment data (about 5 s of speech for each enrolled speaker) and test data (2–8 s) within a text-independent mode. We present a simple and yet powerful decision rule to accept or reject speakers using a discriminative vector in the decision score space, together with statistical hypothesis testing based on the conventional likelihood ratio test. Discriminative training is introduced to further improve system performance for both decision techniques, by employing minimum classification error and minimum verification error frameworks. Experiments are performed using three separate corpora. Using the YOHO speaker recognition database, the alternative decision rule achieves measurable improvement over the likelihood ratio test, and discriminative training consistently enhances overall system performance with relative improvements ranging from 11.26%–28.68%. A further extended evaluation using the TIMIT (CORPUS1) and actual noisy aircraft communications data (CORPUS2) shows measurable improvement over the traditional MAP based scheme using the likelihood ratio test (MAP-LRT), with average EERs of 9%–23% for TIMIT and 13%–32% for noisy aircraft communications. The results confirm that an effective in-set/out-of-set speaker recognition system can be formulated using discriminative training for rapid tagging of input speakers from limited training and test data sizes.**

***Index Terms*—Decision score space, discriminative training, in-set/out-of-set, limited training data, minimum classification error, minimum verification error, speaker recognition.**

I. Introduction

IN MANY speech systems, it is desirable to be able to detect and track the presence of a group of speakers. For example, in spoken document retrieval (SDR), it is useful to identify if a speaker within a group appears repeatedly across an audio stream such as tracking TV anchors, and separate these speakers from those being interviewed. Other speech systems that can benefit from in-set/out-of-set recognition are dialog systems, speech communications systems, speaker clustering for acoustic model adaptation, and applications that allow security and proper access to private information for a specific group of people. Various government agencies use voice authentication for security purposes, such as ensuring that only authorized users have access to computer files or building access [31]. However, traditional open-set speaker recognition typically employ either moderate to large amounts of enrollment data for text-independent testing conditions, or are text dependent for limited enrollment/testing applications.

The general goal of speaker recognition is to require a machine to automatically recognize a speaker from his/her voice [5]. In speaker recognition applications, two well-known problems have received much research attention in the speech community: closed-set speaker identification [12] and speaker verification [19]. In the speaker verification problem, the objective is to decide whether to accept or to reject the identity of a claimed speaker requiring a binary yes/no answer. In the closed-set speaker identification problem, the objective is to classify an unknown observation $\mathbf{X}$ into one of a pre-defined set of enrolled speakers (or classes) $\{S_n | n = 1, \ldots, N\}$, where N is a total number of enrolled speakers. If $\mathbf{X}$ is actually from one of the pre-defined classes, the optimal class decision S^* that maximizes the *a posteriori* probability is given by

$$S^* = \arg\max_{1 \le n \le N} p(S_n|\mathbf{X}) = \arg\max_{1 \le n \le N} p(\mathbf{X}|S_n) \cdot p(S_n) \quad (1)$$

where the conditional probability $p(\mathbf{X}|S_n)$ and the *a priori* probability $p(S_n)$ are assumed known. However, in some applications or practical situations, $\mathbf{X}$ may belong to none of the pre-defined classes, namely an *outlier* or *out-of-set* speaker. In particular, by relaxing the *closed-world* assumption[1] to the closed-set speaker identification problem, one can merge the two tasks, closed-set speaker identification and speaker verification, for what is called open-set speaker identification (OSI). One can adopt two statistical stages for the OSI problem: first classify $\mathbf{X}$ into the most likely class, (e.g., $S_n (1 \le n \le N)$), then verify whether $\mathbf{X}$ actually comes from the class S_n (accept) or if $\mathbf{X}$ is an outlier (reject). Therefore, the objective of the open-set problem is to classify an unknown observation into one of the $(N + 1)$ classes: with N pre-defined classes and a single *none-of-them* class.

Recently, several research studies have reported promising approaches for the open-set problem. Jiang and Deng [15]

Manuscript received November 4, 2004; revised May 30, 2005. This work was supported by the U.S. Air Force Research Laboratory, Rome, NY, under Contract F30602-03-1-0110. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Timothy J. Hazen.

The authors are with the Center for Robust Speech Systems, University of Texas at Dallas, Richardson, TX 75083-0688 USA (e-mail: angkitit@utdallas.edu; john.hansen@utdallas.edu).

Color versions of Figs. 3–8 are available online at http://ieeexplore.ieee.org.
Digital Object Identifier 10.1109/TASL.2006.881689

[1]For the closed-world assumption, $\mathbf{X}$ always belongs to one of the pre-defined classes.

proposed a Bayesian approach (the Bayes factor) for Gaussian mixture models (GMMs), where the Viterbi approximation is adopted in the computation of the joint Bayesian predictive densities for the outlier verification problem in pattern recognition. Gong [13] focused on the in-vehicle open-set speaker recognition task, using a new decision procedure, which treats the scores as the outcome of a Gaussian mixture distribution using a hidden Markov model (HMM)-based recognizer where the mean and covariance parameters are modeled as polynomial functions of the noise level. Deng and Hu [7] proposed a system which uses vector quantization (VQ) for speaker modeling, and a support vector machine (SVM) binary classifier to form a set-score and a vector distance measurement for each reference speaker to either accept/reject the speaker. Murthy *et al.* [22] employed a cepstral slope feature together with variance transformations within a GMM framework. Angkititrakul and Hansen [2] investigated an alternative hypothesis testing method using the speaker's neighborhood information in a speaker model space. Finally, various score normalization techniques for text-independent open-set speaker identification were investigated by Sivakumaran *et al.* [29].

Some related research to the open-set speaker recognition problem include utterance verification [23], topic verification [18], and information retrieval (i.e., National Gallery of the Spoken Word [14]). Utterance Verification attempts to reject or accept part or all of an utterance based on a computed confidence score. This method also attempts to reject erroneous but valid keyword strings. Utterance verification is formulated as a statistical hypothesis test where the task is to evaluate the null hypothesis that a given keyword or a set of keywords exists within a segment of speech against the alternative hypothesis that such a keyword or keyword set does not exist, or is incorrectly classified, within that speech segment. The goal of topic verification is to make a decision as to whether a document truly belongs to a particular topic of interest. Information retrieval (IR) typically is focused on the scenario of identifying items or documents within a large collection that best match a query provided by a user. These research areas share many of the same classification goals required for identifying members belonging to a group like the open-set problem.

In this paper, we consider the generalized problem of open-set speaker recognition, namely in-set/out-of-set speaker recognition. The objective of in-set/out-of-set speaker recognition is to make a decision regarding the identity claimed by a speaker as to whether to accept or reject as being a legitimate member of a group of enrolled speakers. Thus, in-set/out-of-set speaker recognition requires a binary decision: is this an *in-set* or *out-of-set* speaker? The challenge for in-set/out-of-set speaker recognition is effective rejection of outliers, while dealing with distinct dimensions of in-set speaker variability such as interspeaker variations at both the segmental and supramental levels [9]. The evaluation is based on two types of error measurements, namely false rejection (FR) and false acceptance (FA). The former type of error occurs when an in-set speaker is rejected as a legitimate member of the group, whereas the latter occurs when a non-legitimate speaker is accepted as a member of the group. Unlike open-set speaker recognition, we count all acceptances within the in-set speaker group as correct classification, even if the particular in-set speaker is not correctly classified within the group. In-set/out-of-set speaker recognition is therefore very close to the problem of outlier rejection in pattern recognition, when the outliers are unknown speakers. Therefore, in-set/out-of-set recognition can be viewed as a special case of open-set speaker recognition where we combine all in-set confusions as successful in-set recognitions. The present problem is, however, significantly different from traditional speaker identification tasks because of the severe limitations in training and test data sizes.

The present study is focused on the text-independent task, with the constraint of limited-enrollment speech (e.g., approximately 5 s of speech for each enrolled speaker) and limited test data 2–8 s. A medium quantity of development speech is employed for building a universal background model (UBM) and decision threshold. Recently, GMMs [25] have been the dominant technique in text-independent speaker recognition because of their high performance when sufficient training and test data sizes are available. The amount of enrollment data required for a GMM speaker recognition system varies according to the application (e.g., for a high-security application, several minutes or more of enrollment speech might be required for each enrolled speaker in order to have a detailed description of the acoustic phoneme space). Unfortunately, the conventional GMM does not perform well if enrollment data is sparse (e.g., several seconds). Moreover, with short-enrollment data, we need to deal with the problem that phonemes may be present in the test but not in the training material. One technique to mitigate the sparseness issue is to create a UBM based on a larger amount of speech data, compared to each individual speaker data, from nontarget or development speakers. With this approach, a speaker-dependent GMM is then created by maximum *a posteriori* (MAP) adaptation of the UBM [26].

In this paper, we propose an integrated system which employs discriminative training to minimize the misclassification and misverification errors at the training stage. The classical minimum classification error (MCE) criterion is employed to minimize the classification error among the in-set speaker models, followed by the minimum verification error (MVE) criterion which aims to minimize the verification error of the in-set speaker models and the background model. Siohan *et al.* [28] and Rosenberg *et al.* [27] were the first to successfully employ MCE and MVE schemes to speaker identification and speaker verification tasks, respectively. In our study, we modified the MVE framework to adapt the background model to minimize the verification error of each in-set group of speakers. That is, instead of having one common UBM for all in-set speaker groups (we assume that as the number of in-set speakers increases, they will need to be clustered into smaller sets of in-set speakers using for example a neighborhood information [2]), we use a different adapted background model for each individual group of in-set speakers. We present a simple and yet powerful decision rule for outlier rejection for in-set/out-of-set speaker recognition by investigating the distribution of discriminant vector in the N-dimensional decision space, which is spanned by the N-target vectors, and compared this performance with the conventional likelihood ratio test (LRT).

This paper is organized as follows. In Section II, we briefly review the objective of open-set speaker identification and in-set/out-of-set speaker recognition. Section III describes acoustic model training, starting with the GMM-UBM and followed by MCE and MVE frameworks. Section IV discusses the outlier rejection criteria applied for in-set/out-of-set speaker recognition using the discriminant vector on the decision space. We apply all methods to data from the YOHO speaker recognition database and report evaluation results in Section V. Section VI reports on an extended set of experimental results from two additional speech databases, namely one with phonetically balanced sentences (CORPUS1) and noisy aircraft communications (CORPUS2). Finally, the conclusion will be given in Section VII.

II. In-Set/Out-of-Set Speaker Recognition

Let us assume that we are given a set of N in-set (enrolled) speakers in a system, with a collection of observations $\mathbf{X}_n$, corresponding to each enrolled speaker S_n, $1 \leq n \leq N$. Let $\mathbf{X}_0$ represent all other observations from the non-enrolled speakers in the development set. Each speaker-dependent statistical model Λ_n, $\{\Lambda_n \in \mathbf{\Lambda}, 1 \leq n \leq N\}$, can be obtained from $\mathbf{X}_n = \{\mathbf{x}_{n1}, \mathbf{x}_{n2}, \ldots, \mathbf{x}_{nT_n}\}$ where T_n denotes the total number of observations that belong to speaker S_n.

If $\mathbf{X}$ denotes the sequence of observation vectors extracted from the test utterance, then the problem of open-set speaker identification requires that we perform the following two steps. In the first stage, called *(closed-set) speaker identification* or *speaker classification*, we first classify $\mathbf{X}$ into one of the most likely in-set speakers Λ^* as

$$\Lambda^* = \arg\max_{1 \leq n \leq N} p(\mathbf{X}|\Lambda_n). \tag{2}$$

In the second stage, called *speaker verification* or *outlier verification*, we verify whether the observation $\mathbf{X}$ truly belongs to Λ^* or not (i.e., accept/reject). In general, this stage is formulated as a problem in statistical hypothesis testing when the *null* hypothesis $\mathbf{H}$ represents the hypothesis that $\mathbf{X}$ really belongs to speaker model Λ^*, against the competitive hypothesis $\mathbf{H}'$, that represents the hypothesis where $\mathbf{X}$ is actually *not* the speaker model Λ^*. If the probabilities of the null and the alternative hypotheses are assumed known, then according to the Neyman–Pearson Lemma, the conventional LRT is optimal [10]

$$\frac{p(\mathbf{X}|\Lambda^*)}{p(\mathbf{X}|\Lambda_0)} \begin{cases} \geq \gamma & : \text{accept } \mathbf{H} \\ < \gamma & : \text{reject } \mathbf{H} \text{ (accept } \mathbf{H}') \end{cases} \tag{3}$$

where γ is a pre-defined threshold, Λ_0 is a competitive or anti-speaker model (e.g., UBM or cohort-speaker models), and $p(\cdot|\cdot)$ is the likelihood given each speaker model Λ. In practice, it is impossible to have a true antimodel for the competitive speaker class, otherwise we could define such a speaker model as one class in the training phase. The conventional strategy assumes another special class, or background model, which is speaker independent, as a universal representative of all other speakers excluding the in-set speakers (that is, a class which includes all possible outliers).

From the context of open-set speaker identification, the problem of in-set/out-of-set speaker recognition can be performed by relaxing the *null* hypothesis $\mathbf{H}$ as $\mathbf{X}$ really belonging to the *in-set* speaker group against the alternative hypothesis $\mathbf{H}'$ that $\mathbf{X}$ belongs to the *out-of-set* speaker group. That is, in-set/out-of-set speaker recognition requires only a binary decision: does $\mathbf{X}$ belong to one of the in-set speakers, or to none of them.

III. Speaker Modeling

A. GMM–UBM

Recently, the GMM employing a UBM with MAP speaker adaptation has become the dominant approach in text-independent speaker recognition [26], [32]. In this section, we briefly review the GMM–UBM system as our baseline system. A speaker-independent model, or UBM, is trained from the non-target speakers using the expectation maximization (EM) algorithm. The probability density function (*pdf*) of an M-component Gaussian for the D-dimensional observation vector $\mathbf{x}_t$, given a particular speaker model Λ_n, is defined as

$$p(\mathbf{x}_t|\Lambda_n) = \sum_{m=1}^{M} \omega_{nm} \mathcal{N}_{nm}(\mathbf{x}_t) \tag{4}$$

where ω_{nm} is the mixture weight of the mth component unimodal Gaussian density $\mathcal{N}_{nm}(\mathbf{x}_t)$, with each parameterized by a mean vector μ_{nm} and covariance matrix $\mathbf{\Sigma}_{nm}$, which is assumed diagonal

$$\mathcal{N}_{nm}(\mathbf{x}_t) = \frac{1}{(2\pi)^{\frac{D}{2}} |\mathbf{\Sigma}_{nm}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}_t - \mu_{nm})^T \mathbf{\Sigma}_{nm}^{-1} (\mathbf{x}_t - \mu_{nm})}. \tag{5}$$

Given that each individual Gaussian *pdf* integrates to unity, then the mixture weight must further satisfy the constraint $\sum_{m=1}^{M} \omega_m = 1$. Consequently, a speaker model can be represented as a set of GMM means, covariance matrices, and mixture weights, denoted as $\Lambda_n = (\boldsymbol{\omega}_{nm}, \boldsymbol{\mu}_{nm}, \mathbf{\Sigma}_{nm})$, for $m = 1, \ldots, M$ and $n = 1, \ldots, N$. For each target speaker, a speaker-dependent GMM can be created by MAP adaptation of the UBM parameters $\{\boldsymbol{\omega}_{0m}, \boldsymbol{\mu}_{0m}, \mathbf{\Sigma}_{0m}\}$ assuming sufficient training data $\mathbf{X}_n = \{\mathbf{x}_{n1}, \mathbf{x}_{n2}, \ldots, \mathbf{x}_{nT_n}\}$. Based on experimental results, the best performance can be achieved using only mean adaptation for each Gaussian component. The mean $\hat{\mu}_{nm}$ of the mth component of the Λ_n is updated via the following formula:

$$\hat{\mu}_{nm} = \frac{\eta_m}{\eta_m + r} E_m(\mathbf{X}_n) + \frac{r}{\eta_m + r} \mu_{nm} \tag{6}$$

where r is a relevance factor which depends on the parameter dimension and controls the balance of adaptation; η_m and $E_m(\mathbf{X}_n)$ can be computed as

$$P(m|\mathbf{x}_{nt}) = \frac{\omega_{nm} \mathcal{N}_{nm}(\mathbf{x}_{nt})}{\sum_{j=1}^{M} \omega_{nj} \mathcal{N}_{nj}(\mathbf{x}_{nt})} \tag{7}$$

$$\eta_m = \sum_{t=1}^{T_n} P(m|\mathbf{x}_{nt}) \tag{8}$$

$$E_m(\mathbf{X}_n) = \frac{1}{\eta_m} \sum_{t=1}^{T_n} P(m|\mathbf{x}_{nt}) \cdot \mathbf{x}_{nt}. \tag{9}$$

The speaker-dependent model obtained from a MAP-adapted UBM provides a tighter coupling between the speaker specific model and the UBM, and helps mitigating the sparseness issue of limited enrollment data for our present formulation. We note that with 5 s of training data per speaker, it is expected that either limited or no data will available for some portions of the acoustic phoneme space. Therefore, discriminative training is suggested in order to better leverage in-set versus out-of-set speaker models.

B. Discriminative Training

If the observations for speaker modeling are distributed according to the assumed statistical model, the optimal training technique is maximum likelihood estimation (MLE). Unfortunately, with sparse training data the true distribution of the speaker acoustic space cannot be modeled reliably with a GMM. Discriminative training methods attempt to minimize the error rate more effectively by utilizing both the correct and alternative categories, and incorporating that information into the training phase. In this section, MCE and MVE discriminative training are applied to estimate the parameters of the hypothesis test so as to minimize the errors from the training data. The MCE discriminative training paradigm has previously been applied successfully for speaker identification [28], and automatic speech recognition [17]. The MVE discriminative training was successfully applied for speaker verification [27], utterance verification [23], [24], and topic verification [18]. Unlike MLE, the goal of discriminative training is to minimize the expected loss function for classification (and verification). We recognize that MCE and MVE have been applied to the general problem of speaker recognition, and emphasize here that because of limited training data, and our interest in recognition of only in-set speakers, that this formulation builds on these previous studies. Here, we apply both discriminative schemes in two stages, which are: 1) minimize classification error between in-set speakers (adapting the in-set speaker models) and 2) minimize verification error between in-set speakers and the background model (adapting the background models).

1) Adapting In-Set Speaker Models: MCE: Let us assume that the discriminant function family contains N discriminant function $\mathcal{L}(\mathbf{X}|\Lambda_n)$, $n = 1, \ldots, N$. The objective of the first adaptation stage is to minimize the classification errors among the in-set speaker models for the training data. First, a misclassification measure is defined for each speaker, which attempts to measure the misclassification distance of an observation sequence spoken by speaker i as $d_i(\mathbf{X}|\mathbf{\Lambda})$

$$d_i(\mathbf{X}|\mathbf{\Lambda}) = -\mathcal{L}(\mathbf{X}|\Lambda_i) + \max_{j \neq i} \mathcal{L}(\mathbf{X}|\Lambda_j), \qquad j = 1, \ldots, N \tag{10}$$

where the anti-discriminant function $\max_{j \neq i} \mathcal{L}(\mathbf{X}|\Lambda_j)$ is the likelihood of the most likely competing class. Here, the discriminant function $\mathcal{L}(\mathbf{X}|\Lambda_i)$ for the correct class Λ_i is defined as the log-likelihood score generated by the GMM Λ_i as follows:

$$\mathcal{L}(\mathbf{X}|\Lambda_i) = \frac{1}{T} \sum_{t=1}^{T} \log \sum_{m=1}^{M} \omega_{im} \mathcal{N}_{im}(\mathbf{x}_t). \tag{11}$$

To transform $d_i(\mathbf{X}|\mathbf{\Lambda})$ into a normalized smooth function, we can use the sigmoid function to embed $d_i(\mathbf{X}|\mathbf{\Lambda})$ into a zero-one loss function $l_i(\mathbf{X};\mathbf{\Lambda})$ as follows:

$$l_i(\mathbf{X};\mathbf{\Lambda}) = \frac{1}{1 + e^{-\alpha d_i(\mathbf{X},\mathbf{\Lambda})}} \tag{12}$$

where α is a slope coefficient of the sigmoid function. When $d_i(\mathbf{X},\mathbf{\Lambda})$ is a large negative number, indicating a correct recognition, the loss function $l_i(\mathbf{X};\mathbf{\Lambda})$ will have a value close to zero, which implies that no loss was incurred. On the other hand, when $d_i(\mathbf{X},\mathbf{\Lambda})$ is a positive number, it leads to a value between zero and one, thus indicating the likelihood of an error. The overall empirical loss associated with the given observations $\mathbf{X}$ is given by

$$l(\mathbf{X};\mathbf{\Lambda}) = \sum_{i=1}^{N} l_i(\mathbf{X};\mathbf{\Lambda}) \delta_i(\mathbf{X}) \tag{13}$$

where $\delta_i(\mathbf{X})$ is a Boolean function which will return 1 if $\mathbf{X}$ is from class i and 0 otherwise. Thus, the objective of the first stage is to minimize the expected loss function defined as

$$L(\mathbf{\Lambda}) = E\left[l(\mathbf{X};\mathbf{\Lambda})\right]. \tag{14}$$

The parameter $\mathbf{\Lambda}$ can be estimated by first choosing an initial estimate $\mathbf{\Lambda}$ and iteratively adapting the set of parameters using gradient probabilistic descent (GPD) according to

$$\mathbf{\Lambda}^{\{\tau+1\}} = \mathbf{\Lambda}^{\{\tau\}} - \varepsilon_{\{\tau\}} \nabla l(\mathbf{X};\mathbf{\Lambda})|_{\mathbf{\Lambda}=\mathbf{\Lambda}^{\{\tau\}}} \tag{15}$$

where $\varepsilon_{\{\tau\}}$ is a learning rate, $\nabla l(.)$ is the gradient of the loss function, and $\{\tau\}$ is the iteration number. The details of the derivation of the GMM parameters using the GPD technique can be found in [16], [21].

2) Adapting the Background Model: MVE: In contrast to the first stage which attempts to minimize the classification error among the in-set speaker models, the objective of the second stage is to minimize the verification error between the in-set and background models. The misverification measure is now defined as

$$d_i(\mathbf{X};\Lambda_{\mathrm{UBM}}) = -\mathcal{G}_i(\mathbf{X};\mathbf{\Lambda}) + \mathcal{L}(\mathbf{X}|\Lambda_{\mathrm{UBM}}) \tag{16}$$

where the discriminant function $\mathcal{G}_i(\mathbf{X};\mathbf{\Lambda})$ is defined as

$$\mathcal{G}_i(\mathbf{X};\mathbf{\Lambda}) = \log \left[\frac{1}{P} \sum_{j=1}^{P} \exp\left[\eta \cdot \mathcal{L}(\mathbf{X}|\Lambda_j)\right] \right]^{\frac{1}{\eta}}, \qquad \eta > 0. \tag{17}$$

Here, $\mathcal{G}_i(\mathbf{X};\mathbf{\Lambda})$ can be viewed as a geometric mean of the likelihoods of the P-best in-set likelihood scores. The underlying goal is to increase the discriminability between the best in-set candidate models and the universal background model. Similar to the misclassification measure, we transform the misverification measure into a smooth and differentiable zero-one loss function of the form in (12). At this stage, only the parameters of the UBM are adapted according to

$$\Lambda_{\mathrm{UBM}}^{\{\tau+1\}} = \Lambda_{\mathrm{UBM}}^{\{\tau\}} - \varepsilon_{\{\tau\}} \nabla l(\mathbf{X};\Lambda_{\mathrm{UBM}})|_{\Lambda_{\mathrm{UBM}}=\Lambda_{\mathrm{UBM}}^{\{\tau\}}} \tag{18}$$

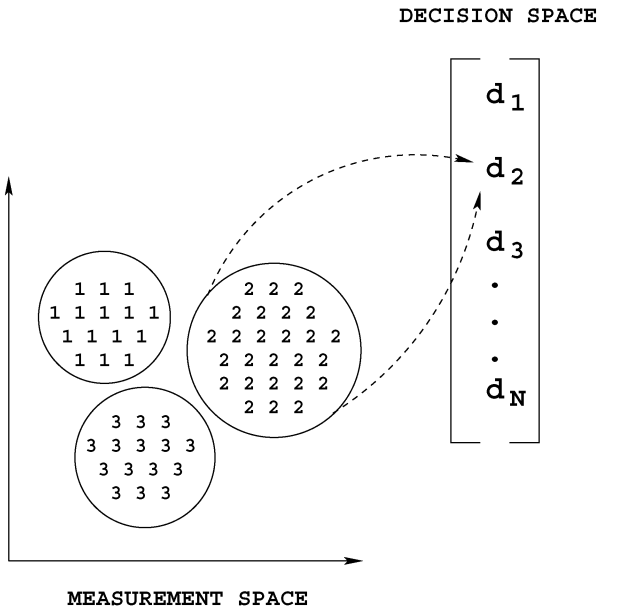


Fig. 1. Mapping from measurement space into a decision space [30].

where, again, $\varepsilon_{\{\tau\}}$ is a learning rate, $\nabla l(.)$ is the gradient of the loss function, and $\{\tau\}$ is the iteration number.

IV. Decision Rules

In general, there are two reasons for rejecting outliers which include: 1) insufficient evidence that the claimant is an in-set speaker and 2) suspicion of the claimant being an out-of-set speaker. To perform in-set/out-of-set speaker recognition in the context of open-set speaker identification, one has to perform statistical hypothesis testing as represented in (3). In practice, the likelihood ratio test is performed using the log-likelihood ratio between the most-likely in-set speaker score and the universal background model score as follows:

$$\mathcal{R}(\mathbf{X}|\Lambda_\kappa) = \mathcal{L}(\mathbf{X}|\Lambda_\kappa) - \mathcal{L}(\mathbf{X}|\Lambda_{\mathrm{UBM}}) \begin{cases} \geq \gamma, & \text{in-set} \\ < \gamma, & \text{out-of-set} \end{cases} \tag{19}$$

where γ is a pre-defined threshold, and κ is the most-likely in-set speaker for the observation sequence $\mathbf{X}$. The decision rule in the context of statistical decision theory is optimal for the case of insufficient evidence if the closed-world assumption holds and if the *a posteriori* probability is known. For comparison, we introduce an alternative decision procedure for outlier rejection, which investigates the distribution of the vector of estimated *a posteriori* probabilities in the decision space [30].

Given a set of N in-set speakers with index n for each speaker $(1 \leq n \leq N)$, the target vector $\mathbf{y}_n$ which corresponds to speaker n, is defined as a unit vector which has only the nth component having the value 1 and all other components are 0. For example, a target vector for the second speaker of a set of three in-set speakers is $[0\ 1\ 0]^T$. Consequently, we let D represent the N-dimensional decision space spanned by the N target vectors $\{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$. In the D space, the discriminant vector $\mathbf{d}$ can be obtained from mapping the measurement space into the decision space using a discriminant function. Fig. 1 illustrates the idea of mapping the input data sequence from the feature space into the decision space. In the present study, we use the *a posteriori* probability as our discriminant function followed by unit normalization as

$$\mathbf{d} = \frac{1}{\sum_{n=1}^{N} \exp\left(\mathcal{L}(\mathbf{X}|\Lambda_n)\right)} \begin{bmatrix} \exp\left(\mathcal{L}(\mathbf{X}|\Lambda_1)\right) \\ \exp\left(\mathcal{L}(\mathbf{X}|\Lambda_2)\right) \\ \exp\left(\mathcal{L}(\mathbf{X}|\Lambda_3)\right) \\ \vdots \\ \exp\left(\mathcal{L}(\mathbf{X}|\Lambda_N)\right) \end{bmatrix}. \tag{20}$$

With a discriminant vector $\mathbf{d}$, the decision concerning the most likely speaker is derived by searching for the maximum component of $\mathbf{d}$ (speaker recognition), and the reject decision can be performed by investigating the distribution of the component values. In our study, we focus on the following two rejection criteria [30]:

$$\text{MAX criterion : out-of-set if } d_{\max} < \beta_{\mathrm{MAX}} \tag{21}$$

$$\text{RAD criterion : out-of-set if } r^2_{\min} > \beta_{\mathrm{RAD}} \tag{22}$$

where β_{MAX} and β_{RAD} are pre-defined thresholds for each criterion, and $r^2_{\min}$ is the minimum among the squared Euclidean distances between the discriminant vector $\mathbf{d}$ and a set of target vectors $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$. The motivation for minimizing the squared Euclidean distance is similar to the idea seen in least mean-square learning techniques. The squared Euclidean distance between $\mathbf{d}$ and the set of N target vectors $\{\mathbf{y}_n\}$, $n = 1, \ldots, N$ can be found using the following relation:

$$r_n^2 = |\mathbf{d} - \mathbf{y}_n|^2 = 1 + |\mathbf{d}|^2 - 2d_n \tag{23}$$

where d_n is the nth component of $\mathbf{d}$. If we assume that all the classes have approximately equal *a priori* probabilities, the MAX criterion will cover the borders between each pair of classes. The MAX criterion rejects patterns whenever the maximum component of the discriminant vector is smaller than a required minimum threshold, while the RAD criterion rejects patterns whenever the squared Euclidean distance between $\mathbf{d}$ and nearest target point exceeds a maximum threshold. For illustration, let us consider a three-dimensional decision space D, where the vertices $\{[1\ 0\ 0]^T, [0\ 1\ 0]^T, [0\ 0\ 1]^T\}$ form a two-dimensional equilateral triangle within D space. Such a triangle plane is shown in Fig. 2. Similarly, for an N-dimensional decision space, the vertices of all target vectors will form an $(N-1)$-dimension decision plane within that decision space. The MAX criterion establishes an island in that region of the decision space which has small likelihoods among the classes, and rejects all patterns that fall inside that island. The RAD criterion establishes isolated circular boundaries for each class, and rejects all patterns that fall outside those regions. The shaded areas in figure show the rejection regions for these two criteria. Such rejection criteria are simple and yet powerful, since they can be employed without training models of the score space, and therefore avoids the score bias problem during training for our limited sized training data set requirements.

V. Experiments

A. Experimental Setup

In order to evaluate the performance of the in-set/out-of-set speaker recognition systems, evaluations were performed using

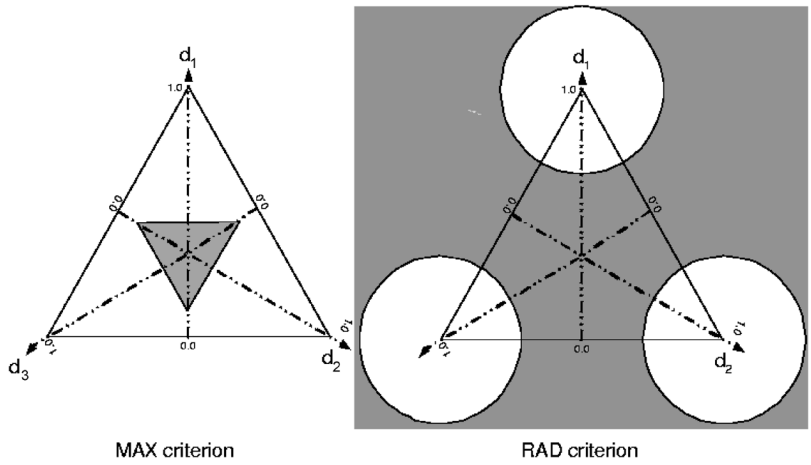


Fig. 2. Two-dimensional decision plane for three-dimensional decision space of the MAX and RAD criteria, where the rejection regions are shaded [30].

three corpora [YOHO, CORPUS1 (TIMIT), and CORPUS2 (noisy, aircraft communications)]. For the primary evaluation, we focus on the YOHO speaker recognition database [4]. YOHO uses prompted combination lock phrases, e.g., "42-87-69," "56-25-56." Each speaker in YOHO has four enrollment sessions with 24 phrases per session. Verification consists of ten sessions with four phrases per session (i.e., total of 40 test phrases for each speaker). The first 60 male speakers was selected as a speaker sample space, while the remaining male speakers from the YOHO was used as the development set. These 60 speakers serve both as in-set speakers and out-of-set speakers depending on the experimental set. That is, when a set of N speakers was selected as in-set speakers, the remaining $(60 - N)$ speakers would act as out-of-set speakers during the test stage. The speakers in YOHO development set were used to trained the UBM. For our experiments, we used the first three sessions consisting of 33 male speakers, which were excluded from the core 60 speakers of evaluation set as described above. In particular, three different in-set speaker populations are considered: 15, 30, and 45. For example, 15 speakers were randomly selected as in-set speakers, with the remaining 45 speakers taking on the role of out-of-set speakers ("15in/45out"). Four different speaker combinations of in-set and out-of-set speakers were randomly selected for each in-set population (i.e., four distinct "15in/45out" groups, four "30in/30out" groups, and four "45in/15out" groups.) Training speech for each speaker was obtained by randomly concatenating three phrases from the first three training sessions, with one phrase from each session. Therefore, training for each in-set speaker consisted of about 5 s of data. We created five experimental sets from five different training speech sets (i.e., using a different set of training data for each speaker to generate each speaker-dependent acoustic model). During the test stage, all 40 phrases (from the test section of the original YOHO database) of these 60 speakers were used for all experiments.

B. Front-End Processing

Feature analysis was performed using 19-dimensional mel-frequency cepstral coefficients (MFCC) on a 30-ms frame basis with a 10-ms skip rate. All speech was pre-emphasized with the filter $(1 - 0.95z^{-1})$, and a Hamming window applied to the result. Additionally, four-dimensional subband spectral gravity center (SSGC) [1] features were appended. SSGCs were computed from four non-overlapped subbands of [0–4000] Hz, where 1000 Hz was allocated per subband. Here, we let the lower and higher filter edges of the mth subband be l_m and h_m, respectively, and its filter shape be represented as $v_m(f)$. The mth subband spectral centroid $\mathcal{C}_m$ can then be defined as follows:

$$\mathcal{C}_m = \frac{\int_{l_m}^{h_m} f v_m(f) P^{\gamma}(f) df}{\int_{l_m}^{h_m} v_m(f) P^{\gamma}(f) df} \tag{24}$$

where $P(f)$ is the power spectrum, and γ is a constant controlling the dynamic range of the power spectrum. For our evaluations here, γ was set to 1.0. Here, we used the SSGCs as a crude approximation of the formant locations. Silence and low-energy speech parts were removed using a general frame energy detection technique. Cepstral mean normalization (CMN) [11] was applied on an utterance-by-utterance basis.

C. GMM–UBM System

A UBM is trained from speakers in the development set (these were excluded from the core 60 speakers of evaluation set), with 32-component Gaussians. GMM training was initiated with vector quantization (VQ) codebooks with several update iterations, and the GMM parameters are consequently adjusted with EM iterations. A single speaker-dependent GMM for each in-set speaker is then estimated from the UBM based on MAP adaptation. The relevant factor [e.g., r in (6)] is set to 16.0 for all experiments. The number of Gaussian components is also fixed to be 32 for all speakers across all experiments.

The verification decision is made based on one of three schemes:

1) [LRT-UBM] the conventional LRT using the UBM as the anti-speaker model;
2) [MAX] normalized maximum component of the estimated discriminant vector;
3) [RAD] the minimum Euclidean distance of the estimated discriminant vector and the real target vector.

TABLE I
COMPARISON OF POOLED EER OF THE BASELINE SYSTEM (WHERE $D^{(N)}$ MEANS N-DIMENSIONAL DECISION SPACE, AND $D^{(5)}$ MEANS FIVE-DIMENSIONAL DECISION SPACE)

Rejection Criteria	EER (%)		
	15 in-set	30 in-set	45 in-set
LRT-UBM	18.38	22.38	24.23
MAX ($D^{(N)}$)	17.28	18.85	18.42
MAX ($D^{(5)}$)	17.32	18.53	18.46
RAD ($D^{(N)}$)	17.28	18.52	18.24
RAD ($D^{(5)}$)	17.54	18.76	18.59

1) Evaluation Results: The pooled equal error rate (EER) is the equal error of the system assuming a constant threshold across all speakers. For our evaluation, we use this EER to measure and compare system performance. This performance measure is a reasonable gauge of system performance since it is difficult in practice to optimize the speaker adapted thresholds needed to obtain the average EER [6]. Table I compares the averaged pooled EER of the baseline system using different decision rules (LRT-UBM, MAX, RAD). For MAX and RAD decision rules, we also compare the results using a discriminant vector of dimension N (i.e., use all in-set likelihood scores) and dimension 5 (use only the top five likelihood scores). We can see that as the population size of in-set speaker group increases, the EER tends to increase as well (i.e., since the diversity of the in-set speakers increases); the baseline LRT-UBM system consistently produced the highest error rate for all tested configurations. The MAX and RAD have EERs that vary from 17.28% to 18.85% as the in-set speaker size increased from 15 to 45, while the LRT-UBM has EERs that vary from 18.38% to 24.23%. The performance of the MAX and RAD criterion are comparable, and using only the top five likelihood scores also gives the performance which is almost as good as using all the in-set likelihood scores (thereby slightly reducing computational complexity).

Figs. 3 and 4 show the histogram distributions of the comparative scores for in-set (right curves) and out-of-set (left curves) speakers. For the LRT with the UBM (Fig. 3), we can see that while the shape of the distributions are slightly different due to the amount of test data, the ranges for the in-set group of speakers are quite consistent (between -2.0 and 4.0), and the out-of-set distribution is moving toward the in-set distribution as the in-set speaker group size increases (i.e., from 15/45 for the top plot, to 45/15 for the bottom plot). For the MAX criterion (Fig. 4), the out-of-set scores are primarily distributed in the lower range (e.g., less than 0.2), while the in-set scores are distributed more uniformly (e.g., between 0.2 to 0.9). Both in-set and out-of-set distributions move toward the left scale as the size of in-set speakers increases (the RAD criterion has a similar score distribution structure to the MAX criterion simply with different scale values.)

The results from this section therefore confirm that using either the MAX or RAD verification decision method improves in-set speaker recognition performance, with greater improvement as the number of in-set speakers increases.

D. System Performance With Discriminative Modeling

Next, we consider performance using discriminative modeling. For each experiment, we use the acoustic models from the LRT-UBM baseline system as the initial models. For each iteration, we reuse the training data from the in-set speakers to adapt the in-set models first, and then adapt the background model. The slope of the sigmoid function α [α from (12)] is set to 0.5, the initial learning rate [$\varepsilon_{\{\tau\}}$ from (15)] is set to 0.8, with a scale down factor of 0.8 applied every three iterations, η [from (17)] is set to 1.0 for the average of all P-best in-set likelihood scores, and the number of P best in-set scores is adapted so that the misverification measure is nonnegative for each observation sequence. For our evaluation, we used a fixed setup for all experiments. In practice, optimal tuning for each in-set group will

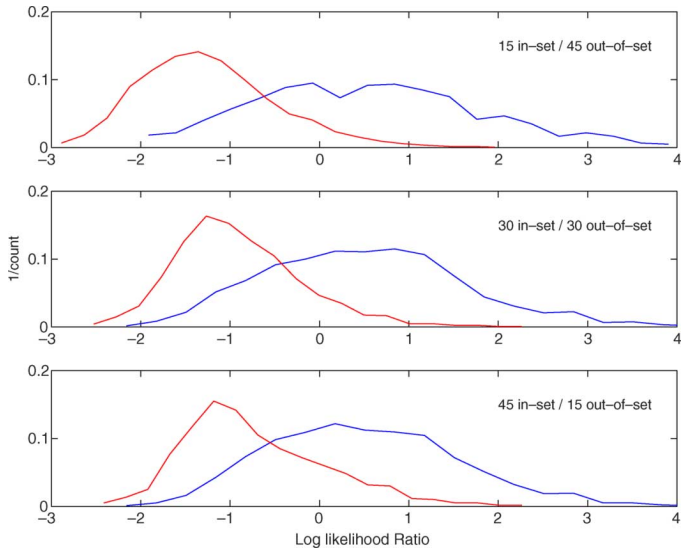


Fig. 3. Histogram showing the distribution of the likelihood ratio scores of various size of in-set speakers (in-set speakers: right curves; out-of-set speakers: left curves).

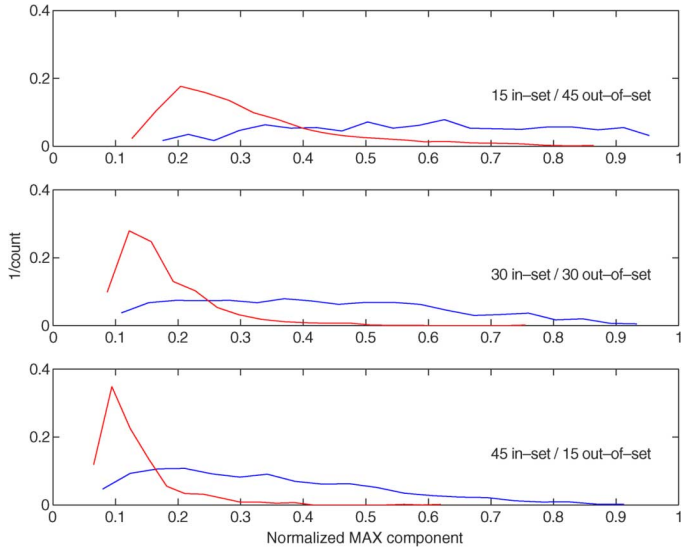


Fig. 4. Histogram showing the distribution of the normalized maximum score component of various size of in-set speakers (in-set speakers: right curves, out-of-set speakers: left curves.

TABLE II
CLOSED-SET SPEAKER IDENTIFICATION

Modeling	Identification accuracy(%)		
	15 in-set	30 in-set	45 in-set
MAP	94.31	91.30	90.13
MAP-DISC	95.00	92.17	91.31

TABLE III
COMPARISON OF POOLED EER USING DISCRIMINATIVE MODELS (WHERE $D^{(N)}$ MEANS N-DIMENSIONAL DECISION SPACE)

Method	EER (%)		
	15 in-set	30 in-set	45 in-set
MAP-LRT	18.38	22.38	24.23
DISC-LRT	18.20	21.80	23.68
MAP-MAX ($D^{(N)}$)	17.28	18.85	18.42
DISC-MAX ($D^{(N)}$)	16.31	17.68	17.45
MAP-RAD ($D^{(N)}$)	17.28	18.52	18.24
DISC-RAD ($D^{(N)}$)	16.43	17.48	17.28

produce higher performance. Only mean and variance parameters are adapted,[2] on a training token by training token basis. Fifteen iterations are applied for each adaptation.

1) Closed-Set Speaker Identification: Although our evaluation is focused on in-set/out-of-set speaker recognition, this section shows results for closed-set speaker identification within the same experimental sets. We see from Table II, that discriminative training, "DISC," consistently improves performance of the baseline system for all in-set speaker sizes, with approximately 0.9% absolute accuracy (or 11.29% relative decrease in error) increase on average.

2) In-Set/Out-of-Set Speaker Recognition: Next, in Table III we compare the pooled EER of the system using discriminative models. We can see that discriminative training methods consistently improve system performance for all decision rules. We note here that the MAX and RAD criterion do not require the background model. From the table, the proposed discriminative scheme achieves slight but consistent improvement over the baseline MAP-LRT method. For comparison, we also illustrate system performance when we change the critical threshold continuously using the detection error tradeoff (DET) curve [20]. Fig. 5 shows an example DET curve of the experiment for the "30in/30out" speaker group. There is measurable improvement over the baseline MAP-LRT system as the false alarm and miss probabilities vary.

E. Open-Set Speaker Identification

In the last experiment with the YOHO database, the evaluation performance is considered in the context of the open-set speaker recognition problem. In this section, the test speech is recognized as `in-set` if it is actually uttered by the hypothesized speaker and `out-of-set` if it is not actually uttered by the hypothesized speaker (Note, that the difference between the evaluations in this section and in the previous section is our definition of the error matrices). Thus, the false acceptance

[2]Our experiments showed that by introducing a small degree of adaptation of the variance parameters in addition to the means, produced improved performance. In this evaluation, we introduced a relative weight for the adaptation of the variance parameters of 0.01, compared to the mean adaptation weight which was 1.0.

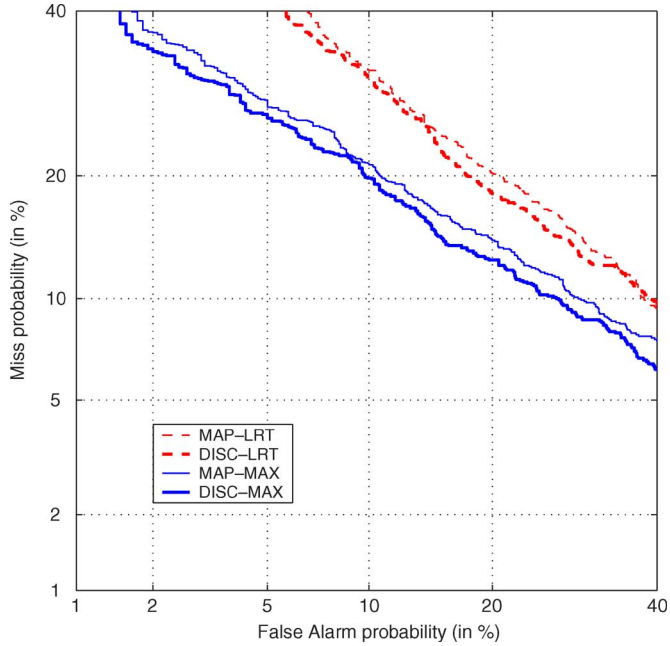


Fig. 5. Example DET curve of the "30in/30out" for different systems.

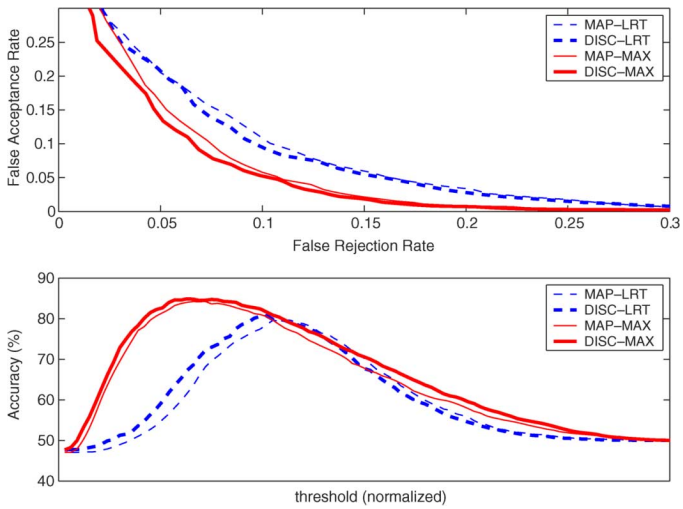


Fig. 6. Top: false acceptance and false rejection curve. Bottom: open-set speaker identification accuracy (%) as threshold varied; for "30in/30out."

(FA) error will include the error of accepting an observation sequence with an incorrect in-set speaker, and the false rejection (FR) error is the same as the previous sections. Using all the test data, we count false acceptance (accept out-of-set speakers and misaccept in-set speakers) and false rejection (reject in-set as out-of-set), and then normalize this value by the total number of test utterances. Using this procedure, Fig. 6 shows the FA–FR plot when we change the critical threshold continuously, using the same test as the DET in Fig. 5. Similar to the DET curve, the plot closer to the bottom left corner shows the better system performance. The lower plot in Fig. 6 shows the open-set speaker identification accuracy (%) as the threshold varies from the minimum score to the maximum score. From this figure, the maximum [and mean] accuracy for MAP–LRT, DISC–LRT, MAP–MAX, and DISC–MAX are 79.71 [59.62], 80.71 [60.15],

84.25 [64.68], and 84.88 [66.15], respectively. We can see that the decision using the MAX criterion outperforms the LRT, and discriminative training further improves system performance for both decision techniques in both FA–FR and accuracy plots.

VI. Extended Evaluation: CORPUS1 and CORPUS2

In this section, we extend our in-set/out-of-set speaker recognition evaluation to consider two additional databases, namely CORPUS1 (TIMIT: clean speech) and CORPUS2 (noisy aircraft communications speech) databases. In particular, CORPUS1 was recorded in a sound booth environment with wide-band frequency content and clean channels, while CORPUS2 was recorded during pilot-to-tower communications with heavy aircraft cockpit noise. The common scenario for all evaluation setup is that the training data for each in-set speaker was extremely limited to approximately 5 s worth of speech, while test data was created for 2, 4, 6, and 8 s worth of speech (of course excluded from the training data). We again employ a framework similar to YOHO, with three partitions consisting of 25%/75%, 50%/50%, and 75%/25% in-set/out-of-set speakers. The purpose for considering these two additional corpora is to verify that the advances seen in YOHO (with low text perplexity using combination digit strings) would carry over to more diverse corpora based on text perplexity and actual noisy conditions.

A. CORPUS1: Clean Speech Corpus (TIMIT)

A set of 60 male speakers were randomly selected as the speaker sample space from the original TIMIT corpus. These 60 speakers serve both as in-set speakers and out-of-set speakers depending on the experimental set. In particular, three different sizes of in-set speakers are considered (e.g., 15, 30, and 45). For example, 15 speakers were randomly selected from the speaker sample space as the in-set speakers, with the remaining 45 speakers taking the role of impostors ("15in/45out"). Similar to other round-robin test procedures, different combinations of in-set and out-of-set speakers were also selected, resulting in four distinct "15in/45out" groups, two distinct "30in/30out" groups, and two (with some overlap) "45in/15out" groups. The training and test speech data of each speaker were randomly selected and concatenated from the original TIMIT database, with no data overlap.

In order to determine a more reliable measure of system performance, another two training and test data sets, similar to the above setup, were also created for the same 60 speakers for comparison. As a result, there are 12 experiments of "15in/45out," six experiments of "30in/30out," and six experiments of "45in/15out." Also, some speakers excluded from the working speaker sample space are used as development data (60 speakers, consisting of approximately 20 min worth of speech).

B. CORPUS2: Noisy Speech Corpus

The second data test set consists of actual speech recorded with aircraft cockpit noise. The transmissions are short in duration and have multiple recording phases (i.e., contains session-to-session variability). Again, our focus is short duration training and test data sizes. Similar to the experimental framework for CORPUS1, a collection of 36 speakers were selected as the speaker sample space. The size of different in-set/out-of-set speaker groups are "9in/27out"(four groups), "18in/18out"(two groups), and "27in/9out" (two groups). The training data was also limited to approximately 5 s worth of speech per speaker, and test data was again created for 2, 4, 6, and 8 s worth of speech. Two sets of training and test for each speaker group were also created for overall averaging of the results. Finally, there are eight experiments of "9in/27out," four experiments of "18in/18out," and four experiments of "27in/9out." Also, some speakers excluded from the working speaker sample space are used as development data (25 speakers, approximately 70 min worth of speech). Finally, we tested a portion of the noisy CORPUS2 data using NIST STNR tool,[3] and found that our data has a STNR value of 19 dB, which is much more noisy than CORPUS1, which has a value of 39 dB.

C. Front-End Processing: CORPUS1 and CORPUS2

In the same manner as the YOHO evaluation, a vector of 19-dimensional MFCCs appended with four-dimensional SSGCs were extracted on a frame basis. For CORPUS1, silence and low-energy speech parts were removed using a general energy detection. For CORPUS2, frames that have a certain number of approximated formant locations that lie within a specified threshold are selected [3].[4] CMN was applied on an utterance-by-utterance basis.

D. Evaluation Results: CORPUS1 and CORPUS2

In this extended evaluation, the number of Gaussian mixtures for all GMMs and UBMs is also fixed to be 32 (note, we continue to use a frame energy threshold to set aside low-energy phonemes which is the reason for 32 mixtures being selected here). Figs. 7 and 8 show plots of average pooled EERs for different test utterance durations of three in-set speaker population sizes, for CORPUS1 and CORPUS2, respectively. As we can see, the EERs of the baseline MAP–LRT tend to degrade as the size of the in-set speaker group increases, and better performance can be achieved as the duration of test utterance increases. For CORPUS1, EERs are in the range of 7%–29% (min-max), with the average spans from 9% to 23% for all four test durations. For CORPUS2, EERs are in the range of 0%–36%, with the average spans ranging from 13% to 32% for all four test durations. The increase in EER for CORPUS2 is expected because of the diversity of the type and level of background noise versus that for the clean CORPUS1. Discriminative training using both DISC–LRT and DISC–MAX systems show consistent improvement over the baseline MAP–LRT system. For CORPUS1, DISC–MAX shows slightly better performance than DISC-LRT on the average; measurable improvement for DISC–MAX over DISC–LRT can be seen for CORPUS2. The results from this extended evaluation therefore have shown that advances for in-set/out-of-set speaker recognition seen for the YOHO corpus

[3][Online]. Available: http://www.nist.gov/speech/index.htm

[4]Although it is considered difficult to achieve the accurate estimation of formant frequencies in noisy conditions, we are only interested in the range of frequencies for our frame selection. Simple root finding of conventional LP analysis is used in our approach [8].

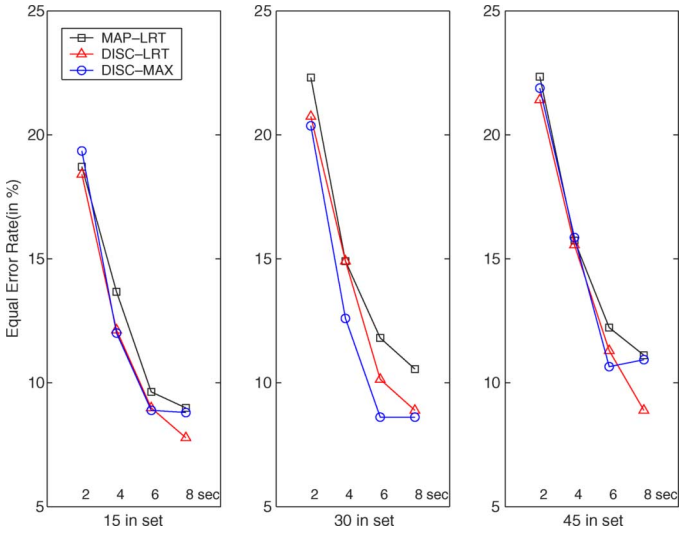


Fig. 7. CORPUS1: comparison of pooled EERs of three different in-set speaker population sizes at 2-, 4-, 6-, and 8-s test utterances.

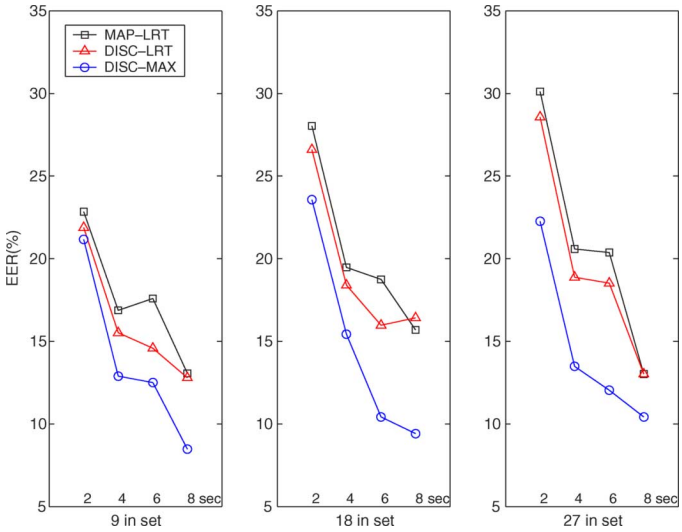


Fig. 8. CORPUS2: comparison of pooled EERs of three different in-set speaker population sizes at 2-, 4-, 6-, and 8-s test utterances.

using MCE adaptation for in-set speaker models, MVE adaptation for the background model, and RAD and MAX decision rules carry over to new in-set/out-of-set speaker recognition scenarios where the amount of training/test data is extremely limited.

VII. Conclusion

In this paper, we have demonstrated an integrated system for text-independent in-set/out-of-set speaker recognition, with extremely short-duration enrollment and test data sizes. The acoustic models were first obtained from a state-of-the-art UBM-GMM employing a MAP adaptation paradigm, followed by a two-step discriminative training scheme that employs MCE and MVE. The discriminative training aims to minimize the classification error among in-set speaker models, and minimize verification errors between in-set speaker models and the background model. We illustrated system performance using the conventional log-likelihood ratio test (LRT) between the most-likely in-set speaker model and the universal background model. We also presented an alternative decision rule for outlier rejection by investigating the distribution of the in-set discriminative score vector itself-without the use of the background model. The experimental results using YOHO, CORPUS1, and CORPUS2 show that the alternative decision rule has lower EERs compared to the log-likelihood ratio test and also is less affected by the population size of the in-set speakers. The results showed that our two-step discriminative training consistently improved system performance for both decision rules using LRT and the in-set discriminative vector. For in-set/out-of-set speaker recognition using extremely limited training and test data sizes, it is clear that a mismatch will exist in the acoustic phoneme space between training/test data sets. The advances made in using MCE and MVE, coupled with the MAX and RAD decision rules, represents an important step forward in addressing the problem of in-set speaker recognition. Future work could consider subspace and clustering of in-set speakers, and also discriminative training on the speaker space. Also, optimal threshold training for real-world applications would also be an area to consider.

References

[1] D. Albesano, R. De Mori, R. Gmello, and F. Mana, "A study on the effect of adding new dimensions to trajectories in the acoustic space," in *Proc. EUROSPEECH/INTERSPEECH*, 1999, pp. 4:1503–4:1506.

[2] P. Angkititrakul and J. H. L. Hansen, "Identifying in-set and out-of-set speakers using neighborhood information," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2004, pp. 1:393–1:396.

[3] K. T. Assaleh and R. J. Mammone, "New LP-derived features for speaker identification," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 630–638, Oct. 1994.

[4] J. P. Campbell, Jr., "Testing with the YOHO CD-ROM voice verification corpus," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1995, pp. 341–344.

[5] ——, "Speaker recognition: A tutorial," *Proc. IEEE*, vol. 85, no. 9, pp. 1437–1462, Sep. 1997.

[6] W. M. Campbell, K. T. Assaleh, and C. C. Broun, "Speaker recognition with polynomial classifiers," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 4, pp. 205–212, May 2002.

[7] J. Deng and Q. Hu, "Open-set text-independent speaker recognition based on set-score pattern classification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Hong Kong, Apr. 2003, pp. II:73–II:76.

[8] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*. New York: IEEE Press, 2001.

[9] G. Doddington, "Speaker recognition-identifying people by their voices," *Proc. IEEE*, vol. 73, no. 11, pp. 1651–1664, Nov. 1985.

[10] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. New York: Academic, 1972.

[11] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. ASSP-29, no. 2, pp. 254–272, Apr. 1981.

[12] H. Gish and M. Schmidt, "Text-independent speaker identification," *IEEE Signal Process. Mag.*, vol. 11, no. 4, pp. 18–32, Oct. 1994.

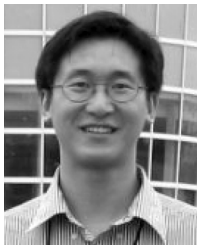
[13] Y. Gong, "Noise-robust open-set speaker recognition using noise dependent Gaussian mixture classifier," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Orlando, FL, May 2002, pp. I:133–I:136.

[14] J. H. L. Hansen, J. Deller, and M. Seadle, "Engineering challenges in the creation of a national gallery of the spoken word: Transcript-free search of audio archives," in *Proc. IEEE and ACM JCDL-2001: Joint Conf. Digital Libraries*, Roanoke, VA, Jun. 24–28, 2001, pp. 235–236.

[15] H. Jiang and L. Deng, "A Bayesian approach to the verification problem: Applications to speaker verification," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 8, pp. 874–884, Nov. 2001.

[16] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Trans. Signal Process.*, vol. 40, no. 1, pp. 3043–3054, Jan. 1992.

[17] B.-H. Juang, W. Chou, and C.-H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 3, pp. 257–265, May 1997.
[18] H.-K. J. Kuo, C.-H. Lee, I. Zitouni, and E. Forler-Lussier, "Minimum verification error training for topic verification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2003, pp. I380–I383.
[19] Q. Li and B.-H. Juang, "Speaker authentication," in *Pattern Recognition in Speech and Language Processing*, ser. Electrical Engineering and Applied Signal Processing. Boca Raton, FL: CRC, 2003, vol. 12.
[20] A. Martin, G. Doddington, T. Kamn, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Sep. 1997, pp. 1895–1898.
[21] E. McDermott, "Discriminative training for speech recognition," Ph.D. dissertation, Graduate School Sci. Eng., Waseda Univ., Tokyo, Japan, 1997.
[22] H. A. Murthy, F. Beaufays, L. P. Heck, and M. Weintraub, "Robust text-independent speaker identification over telephone channels," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 5, pp. 554–568, Sep. 1999.
[23] M. G. Rahim, C.-H. Lee, and B.-H. Juang, "Discriminative utterance verification for connected digits recognition," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 3, pp. 266–277, May 1997.
[24] M. G. Rahim and C.-H. Lee, "String-based minimum verification error (SB-MVE) training for speech recognition," *Comput. Speech Lang.*, vol. 11, pp. 147–160, 1997.
[25] D. A. Reynolds, "Experimental evaluation of features for robust speaker identification," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 639–643, Oct. 1994.
[26] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process.*, vol. 10, no. 1–3, pp. 19–41, Jan./Apr./Jul. 2001.
[27] A. E. Rosenberg, O. Siohan, and S. Parthasarathy, "Speaker verification using minimum verification error training," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1998, vol. 1, pp. 105–108.
[28] O. Siohan, A. E. Rosenberg, and S. Parthasarathy, "Speaker identification using minimum classification error training," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Sep. 1998, pp. 109–112.
[29] P. Sivakumaran, J. Fortuna, and A. M. Ariyaeeinia, "Score normalization applied to open-set, text-independent speaker identification," in *Proc. EUROSPEECH/INTERSPEECH*, Geneva, Switzerland, Sep. 2003, pp. 2669–2672.
[30] J. Schürmann, *Pattern Classification: A Unified View of Statistical and Neural Approaches*. New York: Wiley, 1996.
[31] S. J. Vaughan-Nichols, "Voice authentication speakers to the marketplace," *Computer*, vol. 37, no. 3, pp. 13–15, Mar. 2004.
[32] B. Xiang and T. Berger, "Efficient text-independent speaker verification with structural Gaussian mixture models and neural network," *IEEE Trans. Speech Audio Process*, vol. 11, no. 5, pp. 447–456, Sep. 2003.



Pongtep Angkititrakul (M'04) was born in Khonkaen, Thailand. He received the B.S. degree in electrical engineering from Chulalongkorn University, Bangkok, Thailand, in 1996 and the M.S. and Ph.D. degrees in electrical engineering from the University of Colorado, Boulder, in 1999 and 2004, respectively.

From 2000 to 2004, he was a Research Assistant in the Robust Speech Processing Group, Center for Spoken Language Research (CSLR), University of Colorado. In February 2006, he joined the Center for Robust Speech Systems (CRSS), University of Texas at Dallas, Richardson, as a Research Associate. His research interests are in the general areas of robust speech/speaker recognition, pattern recognition, data mining, human–machine interaction, and speech processing.

John H. L. Hansen (S'81–M'82–SM'93–F'06) received the B.S. degree in electrical engineering from Rutgers University, New Brunswick, NJ, in 1982 and the Ph.D. and M.S. degrees in electrical engineering from the Georgia Institute of Technology, Atlanta, in 1988 and 1983, respectively.

He joined the Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas (UTD), Richardson, in the fall of 2005, where he is Professor and Department Chairman of Electrical Engineering, and holds the Distinguished University Chair in Telecommunications Engineering. He also holds a joint appointment in the School of Brain and Behavioral Sciences (Speech and Hearing). At UTD, he established the Center for Robust Speech Systems (CRSS) which is part of the Human Language Technology Research Institute. Previously, he served as Department Chairman and Professor in the Department of Speech, Language, and Hearing Sciences (SLHS), and Professor in the Department of Electrical and Computer Engineering, University of Colorado, Boulder (1998–2005), where he cofounded the Center for Spoken Language Research. In 1988, he established the Robust Speech Processing Laboratory (RSPL) and continues to direct research activities in CRSS at UTD. His research interests span the areas of digital speech processing, analysis and modeling of speech and speaker traits, speech enhancement, feature estimation in noise, robust speech recognition with emphasis on spoken document retrieval, and in-vehicle interactive systems for hands-free human–computer interaction. He has supervised 36 (18 Ph.D., 18 M.S.) thesis candidates. He is author/coauthor of 222 journal and conference papers in the field of speech processing and communications, coauthor of the textbook *Discrete-Time Processing of Speech Signals*, (IEEE Press, 2000), coeditor of *DSP for In-Vehicle and Mobile Systems* (Springer, 2004–Vol. 1, 2006–Vol. 2), and lead author of the report "The Impact of Speech Under "Stress" on Military Speech Technology," (NATO RTO-TR-10, 2000).

Prof. Hansen is serving as the IEEE Signal Processing Society Distinguished Lecturer for 2005/2006, member of the IEEE Signal Processing Society Speech Technical Committee and Educational Technical Committee, and has served as Technical Advisor to the U.S. delegation for NATO (IST/TG-01). He was Associate Editor for the IEEE TRANSACTIONS SPEECH AND AUDIO PROCESSING (1992–1999), Associate Editor for the IEEE SIGNAL PROCESSING LETTERS (1998–2000), and Editorial Board Member for the IEEE SIGNAL PROCESSING MAGAZINE (2001–2003). He has also served as a Guest Editor of the October 1994 Special Issue on Robust Speech Recognition for the IEEE TRANSACTIONS SPEECH AND AUDIO PROCESSING. He has served on the Speech Communications Technical Committee for the Acoustical Society of America (2000–2003), and is serving as a member of the International Speech Communications Association (ISCA) Advisory Council (2004–2007). He was recipient of the 2005 University of Colorado Teacher Recognition Award as voted by the student body. He also organized and served as General Chair for ICSLP-2002: International Conference on Spoken Language Processing, September 16–20, 2002, and will serve as Technical Program Chair for the IEEE ICASSP-2010.