In-Set/Out-of-Set Speaker Recognition Under Sparse Enrollment

Vinod Prakash, *Student Member, IEEE*, and John H. L. Hansen, *Fellow, IEEE*

***Abstract*—In this paper, the problem of identifying in-set versus out-of-set speakers using extremely limited enrollment data is addressed. The recognition objective is to form a binary decision regarding an input speaker as being a legitimate member of a set of enrolled speakers or not. Here, the emphasis is on low enrollment (about 5 sec of speech for each enrolled speaker) and test data durations (2–8 sec), in a text-independent scenario. In order to overcome the limited enrollment, data from speakers that are acoustically close to a given in-set speaker are used to form an informative prior (base model) for speaker adaptation. Score normalization for in-set systems is addressed, and the difficulty of using conventional score normalization schemes for in-set speaker recognition is highlighted. Distribution scaling based score normalization techniques are developed specifically for the in-set/out-of-set problem and compared against existing score normalization schemes used in open-set speaker recognition. Experiments are performed using the following three separate corpora: 1) Noise-free TIMIT; 2) Noisy in-vehicle CU-Move; and 3) the NIST-SRE-2006 database. Experimental results show a consistent increase in system performance for the proposed techniques.**

***Index Terms*—binary classification, cohort speakers, in-set/out-of-set, in-vehicle CU-move, limited training data, NIST-SRE, score normalization, speaker recognition.**

I. Introduction

IN MANY speech analysis 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) [1], it is useful to identify if a speaker within a group that appears repeatedly across an audio stream such as tracking TV anchors, and separate these speakers from those being interviewed. Other speech-based 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.

Manuscript received February 15, 2007; revised May 26, 2007. This work was supported by RADC under Contract FA8750-05-C-0029 and by University of Texas at Dallas under project EMMITT. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jean-François Bonastre.

V. Prakash was with the Center for Robust Speech Systems (CRSS), University of Texas at Dallas, Richardson, TX 75083-0688 USA. He is with Microsoft Corporation, Redmond, WA 98052 USA (e-mail: vinod.prakash@colorado.edu).

J. H. L. Hansen is with the Center for Robust Speech Systems (CRSS), Department of Electrical Engineering, University of Texas at Dallas, Richardson, TX 75083-0688 USA (e-mail: john.hansen@utdallas.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2007.902058

Speaker recognition is focused on characterizing subjects on the basis of their voice [2]. For in-set/out-of-set speaker recognition problems, the enrollment data is obtained from each individual in a group of speakers, referred to as the in-set group. When presented with an unknown test utterance, the recognizer is required to produce a binary decision as to whether the test utterance came from a speaker belonging to the in-set group or not. This problem is a simplification of the speaker identification open-set case where the recognizer is required to identify a specific speaker within the in-set group or declare that the test utterance is from an out-of-set (i.e., unknown) speaker. Speaker verification is a special case of the in-set problem, where the size of the in-set group is one (in this context, an out-of-set speaker is referred to as an impostor). The Open-set Recognition problem is also referred to as multitarget detection [3] which has been considered in the area of biometrics [4], where in this context the in-set system is referred to as a stack detector. Previous studies in the area of in-set/out-of-set speaker recognition have focused on using discriminative training/scoring [5], clustering [6] or neighborhood information [7] among the in-set speakers.

Speaker recognition systems typically construct statistical models for the enrolled speakers. When statistical models are used to construct representations for the speakers, it has been observed that the raw likelihood scores are not very reliable during the decision process [8]–[10]. If the decision rule for the recognizer depends on the individual speaker models, and out-of-set speaker models, then this process is referred to as score normalization [11]. Two main approaches have emerged to model out-of-set speakers [12], based on the *world model* and *cohort model* schemes. The *world model* [also referred to as a background model or a universal background model (UBM)] is constructed by pooling together data from a large number of potential out-of-set speakers. Alternatively, in the *cohort model* approach, a set of potential impostors is created for each enrolled speaker and score normalization is performed using a statistic of the likelihood scores of these impostors.

In this paper, the problem of in-set speaker recognition is addressed with the constraints of low enrollment (5 s) and test material (2–8 s) and in-set group sizes ranging from 15–45 speakers. When the enrollment data is this low, it is expected that the phoneme coverage for a speaker will be incomplete, resulting in what we call "acoustic holes" in the speaker's model space. When a test token for an enrolled speaker contains phonetic content that was not seen during training, it results in a low likelihood score, and possibly a wrong decision for that speaker (i.e., phonemes seen in the test set but not in the training data for the same speaker cause the speaker model to be rejected). The adapted GMM approach [13] alleviates this problem to a certain extent. Here, the speaker

models are derived by adaptation from a world model. Unseen test data has a similar impact on both the speaker and world models resulting in what is expected to be cancellation of the influence of such data. Since maximum *a posteriori* (MAP) adaptation does not interpolate discriminative information, the adapted-GMM/UBM setup does not score in favor of an in-set speaker when unseen test data from that in-set speaker arrives. The preceding analysis also applies to the case when unseen test data comes from an impostor, because of the limited data available for speaker model adaptation, scores of these impostor test tokens will be comparable to that of the background model, and such impostors are not decisively rejected.

So far, almost all studies involving cohort speakers have used an enrolled speaker's cohorts to model potential impostors. In this paper, we propose to use the cohorts to fill the acoustic holes for the in-set speaker's training space. An initial approach for cohort modeling was shown in [14]. An example for speaker identification was also given in [15] and [16], where, information from cohorts are utilized by merging cohort models, while in our work we utilize information from the cohorts at the feature level.

In-set speaker recognition is generally formulated as a two stage process. Given a test utterance, in the first stage, referred to as *closed-set identification*, the most likely in-set speaker is selected. In the second stage, referred to as the *outlier rejection* stage, a decision is made whether the test utterance could have originated from this most likely in-set speaker or from an impostor. These decisions determine the tradeoff between the two types of error every nonperfect binary classification system can make: *false acceptance* (FA) of an impostor or *false rejection* (FR) of an enrolled speaker. Since a common scale is used to decide in favor of all hypothesized in-set speakers, it is necessary to transform the test statistic (which is most commonly the likelihood score) for all in-set speakers to a common scale. Speaker verification systems face a related issue if, during the evaluations, a common threshold is assumed for all hypothesized speakers. Modeling of the impostor score distributions by a single Gaussian (popularly referred to as Z-norm) [10], numerical simulations [17], and more recently by GMMs [18], have been attempted. While there has been work on application of these schemes to the open-set problem [19]–[22], these have not specifically considered solutions from the *system* perspective. In this paper, we adapt score distribution scaling strategies developed for speaker verification applications to the in-set/out-of-set problem.

This paper is organized as follows. In Section II, we cover the objective formulation of the open-set speaker identification and in-set/out-of-set speaker recognition and provide a brief overview of the baseline system. Section III describes the algorithm used to share information from the cohort speakers. Section IV describes the need for score-normalization for in-set systems and the proposed techniques. We apply the cohort-model-based scheme to data from the TIMIT and CU-Move databases and report evaluation results in Section V. This section also contains experimental results of speaker verification and in-set specific experiments performed on the NIST-SRE database. Section VI provides a discussion and analysis of the results. Finally, conclusions will be given in Section VII.

II. Objective Formulation

Let us assume that a set of N in-set (enrolled) speakers are given for the in-set 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

$$Y_n = p(\mathbf{X}|\Lambda_n) \tag{1}$$

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

$$Y^* = \max_{1\leq n\leq N} p(\mathbf{X}|\Lambda_n). \tag{3}$$

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 likelihood ratio test (LRT) is optimal [23] (in terms of correct detection for a specified false alarm)

$$\frac{p(\mathbf{X}|\Lambda^*)}{p(\mathbf{X}|\Lambda_0)} \begin{cases} \geq \gamma & : \text{accept } \mathbf{H} \\ < \gamma & : \text{reject } \mathbf{H} \ (\text{accept } \mathbf{H}') \end{cases} \tag{4}$$

where γ is a predefined 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. 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. For the in-set (or open-set) case depending on the specific scoring strategy used, the above equation changes. For instance, for the unconstrained cohort normalization (UCN [19] also called DIFF rule [24]) and MAX rule [5], [24], the denominator $p(\mathbf{X}|\Lambda_0)$ in the above equation changes to

$$\text{UCN rule}: p(\mathbf{X}|\Lambda_0) = \max_{1\leq n\leq N, \Lambda_n \neq \Lambda^*} p(\mathbf{X}|\Lambda_n) \tag{5}$$

$$\text{MAX rule}: p(\mathbf{X}|\Lambda_0) = \sum_{n=1}^{N} p(\mathbf{X}|\Lambda_n). \tag{6}$$

A. Baseline GMM-UBM System

The Gaussian mixture model (GMM) employing a UBM with MAP speaker adaptation is the dominant approach in text-independent speaker recognition [13]. In this section, we briefly review the GMM–UBM system as our baseline system. A speaker-independent model, or UBM, is trained from the nontarget speakers using the expectation maximization (EM) algorithm. Every speaker model is represented using a GMM denoted as $\Lambda_n = (\omega_{nm}, \boldsymbol{\mu}_{nm}, \boldsymbol{\Sigma}_{nm})$, for $m = 1, \ldots, M$ and $n = 1, \ldots, N$. Where ω_{nm} is the mixture weight of the mth component unimodal Gaussian density $\mathcal{N}_{nm}(\mathbf{x}_t)$, with each one parameterized by a mean vector $\boldsymbol{\mu}_{nm}$ and covariance matrix $\boldsymbol{\Sigma}_{nm}$, which is assumed diagonal.

For each target speaker, a speaker-dependent GMM can be created by MAP adaptation of the UBM parameters $\{\omega_{0m}, \boldsymbol{\mu}_{0m}, \boldsymbol{\Sigma}_{0m}\}$ 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 $\hat{\boldsymbol{\mu}}_{nm}$ of the mth component of the Λ_n is updated via the following formula:

$$\hat{\boldsymbol{\mu}}_{nm} = \frac{\eta_m}{\eta_m + r_m} E_m(\mathbf{X}_n) + \frac{r_m}{\eta_m + r_m} \boldsymbol{\mu}_{nm} \tag{7}$$

where η_m and $E_m(\mathbf{X}_n)$ can be computed as

$$P(m|\mathbf{x}_{nt}) = \frac{\omega_{nm}\mathcal{N}_{nm}(\mathbf{x}_{nt})}{\sum_{j=1}^{M} \omega_{nj}\mathcal{N}_{nj}(\mathbf{x}_{nt})} \tag{8}$$

$$\eta_m = \sum_{t=1}^{T_n} P(m|\mathbf{x}_{nt}) \tag{9}$$

$$E_m(\mathbf{X}_n) = \frac{1}{\eta_m} \sum_{t=1}^{T_n} P(m|\mathbf{x}_{nt}) \cdot \mathbf{x}_{nt}. \tag{10}$$

Here, r_m is a mixture-specific "relevance factor" that controls the balance of adaptation between the UBM parameters and speaker-specific training observations. This is typically based on the speech frame occurring within a particular pdf. In practice, a global relevance factor is chosen for all mixture components. (We retain a global relevance factor for all the baseline settings in our experiments.)

The speaker-dependent model obtained from a MAP-adapted UBM provides a tighter coupling between the speaker specific model and the UBM, and helps in mitigating the sparseness issue of limited enrollment data to some extent.

III. Cohort-Based Speaker Models

A. Motivation

We begin with the assumption that data can be shared between speakers with acoustic similarity and that such speakers produce many phonemes in a similar way. Based on this, the idea behind the proposed scheme is that, if, for each in-set speaker, a model is built by pooling data from "acoustically close" speakers, that is from that in-set speaker's cohort set, and then if this model were to be MAP adapted using the limited enrollment data, the resulting GMM (which will have no more than 64 components) should be far more representative of the speaker than MAP adaptation from a general UBM which would have a much larger number of pdf components, since a large pool of speakers are used in its construction. In addition, if more development data is available for speakers in the cohort set, we are more likely to be able to fill in the acoustic holes in the training space when only 5 s of data is available for the in-set speaker.

Section III-C gives heuristic details about expected performance of the proposed scheme for different categories of in-set and out-of-set speakers and varying overlap between train and test phonemes. It is to be noted that the proposed method is applicable only in cases where "casual" impostors are expected as out-of-set speakers. If someone is deliberately trying to break into the system, the method fails since by using cohorts as a base model for the in-set speaker we make the system more vulnerable to attack by speakers sounding similar to a given in-set speaker.

B. Steps for Cohort Model Construction

The procedure followed to construct a model for the in-set speaker n, $1 \leq n \leq N$, is as follows:

- For each development speaker i, construct a $\mathrm{GMM}(\Lambda_i^{dev})$ using the training data for that development speaker, for $1 \leq i \leq N_{\mathrm{dev}}$.
- Score each of the above models using the training data $\mathbf{X}_n$ for the in-set speaker:

$$S_i = p(\mathbf{X}_n|\Lambda_i^{dev}), 1 \leq i \leq N_{\mathrm{dev}}. \tag{11}$$

- Sort the scores S_i and pick the top N_{cohort} speakers corresponding to the top N_{cohort} scoring models. N_{cohort} ($\ll N_{\mathrm{dev}}$) is the number of cohorts that are used to fill the acoustic holes for in-set speaker n. These speakers form the cohort set $\Omega_n^{\mathrm{cohort}}$ for this in-set speaker.
- Pool together the data of the selected cohorts and construct a cohort GMM as $\Lambda_n^{\mathrm{cohort}}$ for in-set speaker n.
- Using $\Lambda_n^{\mathrm{cohort}}$ as an initial model for the mean, covariance, and mixture weights, in (7) obtain the in-set speaker model Λ_n.

If Ω_{dev} is the set of all development speakers, construct the set

$$\Omega_{\mathrm{out}} = \Omega_{\mathrm{dev}} - \bigcup_{1 \leq n \leq N} \Omega_n^{\mathrm{cohort}}. \tag{12}$$

Data from a randomly chosen subset of speakers from Ω_{out} is pooled to construct a model for the out-of-set speakers. This model is used for score normalization in both the baseline and proposed systems. For some of the experiments, we considered setting of mixture specific relevance factors. In our experience, for the given task we get consistently better performance by using mixture-specific relevance factors rather than a global relevance factor setting. This is especially true when we vary the size of the cohort set below ten speakers. The mixture-specific relevance factors are obtained as

$$r_m = \gamma * \sum_{t=1}^{T_n} P(m|\mathbf{x}_{0t}) \tag{13}$$

where γ is the inverse of the number of utterances used, and $\mathbf{x}_{0t}$ denote the frames that went into constructing the cohort GMM. The setting of the relevance factors in this fashion is identical

TABLE I
EXPECTED DECISIONS OF IDEAL(SYSTEM-0), BASELINE(SYSTEM-1), AND PROPOSED (SYSTEM-2) SYSTEMS FOR VARIOUS OUT-OF-SET SPEAKERS AND DIFFERING OVERLAP BETWEEN TRAIN AND TEST PHONES. S: SHEEP, W: WOLF, X: UNKNOWN

Speaker	Overlap	System-0	System-1	System-2
In-S	Yes	Accept	Accept	Accept
In-S	No	Accept	X	Accept
Out-S	Yes	Reject	Reject	Reject
Out-S	No	Reject	X	Reject
Out-W	Yes	Reject	Accept	Accept
Out-W	No	Reject	X	Accept

to what was done in [25, Eq. (35)]. The main motivation behind this choice is that when the number of cohorts available for construction of the base GMM for an in-set speaker is low, not all the components in the GMM are modeled with the same level of accuracy. So instead of choosing a constant prior relevance factor for all mixtures, choosing mixture-specific relevance factors depending on the observation frequency of the mixtures results in a better choice.

C. Expected Performance

Borrowing terminology from [26], Table I summarizes (heuristically) the expected behavior of the baseline (a conventional GMM–UBM scheme) and the proposed cohort model scheme for various test scenarios. The in-set speaker is taken to be a "sheep," a default speaker type who dominates the population and for whom systems perform nominally well. The out-of-set speakers are taken to be either "sheep" or "wolves" (speakers who are particularly successful at imitating other speakers).

From Table I, the baseline and the proposed systems are expected to perform well if the out-of-set speakers are sheep and there is overlap between the train and test phonemes. When there is no phoneme overlap, the behavior of the baseline system is unpredictable (noted as X in the table where neither the Accept nor the Reject hypothesis is consistently favored). Assuming that the proposed algorithm System-2 successfully fills the acoustic holes in the training space, under the no-overlap condition, correct (incorrect) results are expected when the out-of-set speakers are sheep (wolves).

IV. SCORE NORMALIZATION

As per (4), a common threshold γ is used during decision making irrespective of which in-set speaker model emerges as the top-scoring speaker model Λ^* in the first stage of open-set detection. Therefore, in order to improve system performance, the impostor (or true) score (or final test statistic) distributions for all in-set speakers need to be transformed so that they are brought into a common scale. Most of the time, impostor score distributions are modeled since there would not be sufficient development data to model the score distribution of the true (i.e., in-set) speakers. A popular way of achieving this in speaker verification systems is to employ the Z-norm [10], for which the impostor score distribution is modeled as a Gaussian random variable, development utterances are used to estimate the mean (μ) and variance (σ^2) of the distribution. For a hypothesized speaker if the test statistic is Y, the transformed test statistic $(Y-\mu)/\sigma$ is used to compare with the predetermined threshold γ. Adaptation of the above scheme to the in-set problem has led to use of the following basic structure for the transformation [19]–[22]:

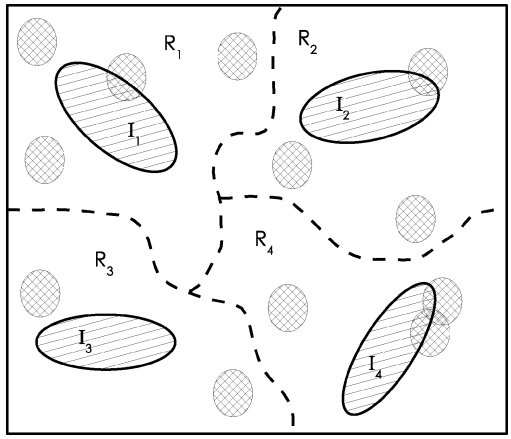

Fig. 1. Abstract illustration of feature space, the circles denote impostors, the ellipses denote in-set speakers, the decision regions shown are where the closed-set identification would be in favor of the respective in-set speakers.

$$Z = \frac{Y(\Lambda^*) - \mu(\Lambda^*)}{\sigma(\Lambda^*)} \tag{14}$$

where Λ^* defined in (2) is the top-scoring in-set speaker model, and $Y(\Lambda_n)$ is the *final* test statistic for the in-set speaker n [left-hand side of (4)]. The mean and variance of the in-set speaker model scores are estimated from a set of development utterances $\mathbf{X_D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t\}$

$$\begin{aligned} \mu(\Lambda_n) &= E[Y_n|\mathbf{x}_i; \mathbf{x}_i \in \mathbf{X_D}] \\ \sigma^2(\Lambda_n) &= var(Y_n|\mathbf{x}_i; \mathbf{x}_i \in \mathbf{X_D}). \end{aligned} \tag{15}$$

The main issue with this approach is illustrated in Fig. 1 for a hypothetical system involving four in-set speakers. The regions R_n denote the locations where, for a given test utterance, the first stage of the open set system decides in favor of in-set speaker S_n. That is, the R_n are the maximum likelihood (ML) decoding regions for the in-set speaker S_n. The circles in the diagram denote development speakers. Taking the example of in-set speaker S_1, if an impostor does not fall in R_1, this impostor will select some other in-set speaker as the top-scoring model in the first stage of open-set detection and so this impostor will not contribute to the impostor score distribution for in-set speaker S_1. Hence, it can be seen that from a system perspective, only the development speakers in a given speaker's R_n will be representative of the impostor score distribution for that specific in-set speaker. This is a direct consequence of selecting the top-scoring model in the first stage of open-set detection, this results in impostors "selecting" models on which they score high, implicitly trying to imitate the in-set speaker. If the other development speakers are used to estimate the score distribution for that in-set speaker, it would be erroneous, since such impostors (or out-of-set) speakers would not follow the pattern of the speakers contributing to making the score of this in-set speaker the highest in the closed-set stage of recognition. Incidentally,

this is the main reason for improved performance of score normalization schemes such as Top-norm [22], where the parameters in (15) are estimated using only the sample values that contribute to the right tail end of the distribution. Essentially, this results in estimating the distribution using those development speakers who cause false alarms for that in-set speaker and who are expected to belong to that in-set speakers ML decoding region.

Hence, from a system perspective, (15) should be modified as follows:

$$\begin{aligned}\mu(\Lambda_n) &= E[Y_n|\mathbf{x}_i; \mathbf{x}_i \in {\mathbf{X_D}}^n] \\ \sigma^2(\Lambda_n) &= \mathrm{var}(Y_n|\mathbf{x}_i; \mathbf{x}_i \in {\mathbf{X_D}}^n)\end{aligned} \tag{16}$$

where

$$\mathbf{x}_i \in {\mathbf{X_D}}^n \text{ if } \Lambda_n = \underset{1\le m\le N}{\arg\max}\ p(\mathbf{x}_i|\Lambda_m).$$

An additional advantage with this scheme is that it can be used with any score normalization technique developed specifically to address the in-set problem (e.g., *RAD or MAX* rule [5], UCN, etc.).[1]

A. Implementation – Score Normalization

In practice, unless a large number of diverse test utterances are used, there would always be a shortage of utterances to estimate the parameters using (16). Generally speaking, if 100 utterances are needed to reliably determine the mean and variance of a Gaussian random variable, for an in-set size of 50 speakers, it would require 5000 utterances to determine the mean and variance parameters for each in-set speaker (in contrast, Z-norm parameters are traditionally estimated using around 100 utterances and Top-norm parameters estimated using typically 1000 utterances). Our solution to this data sparseness issue is to estimate the mean and variance on a "smoothed distribution" by performing a MAP estimation of the parameters using the empirical Bayes technique in place of direct ML estimation.

A joint conjugate prior for the mean and variance parameters of a Gaussian distribution is chosen. The joint conjugate prior is a normal-Gamma distribution [27]. The MAP estimation formulas for the mean and variance parameters are [27], [28]

$$\begin{aligned}\mu_{\mathrm{MAP}} &= \frac{n\bar{y}}{n+r} + \frac{r\mu}{n+r} \\ \sigma^2_{\mathrm{MAP}} &= \frac{\hat{\beta}}{\hat{\alpha}}\end{aligned} \tag{17}$$

where

$$\hat{\alpha} = \alpha + \frac{n}{2}$$

and

$$\hat{\beta} = \beta + \frac{n}{2}S_y^2 + \frac{nr(\bar{y}-\mu)^2}{2(r+n)}$$

where n is the number of samples used in the estimation, $\bar{y}$ is the sample mean, and S_y^2 is the sample variance.

The prior parameters μ, α, β, r needed in the above equations are computed by running all development samples through the in-set recognition system and constructing a one-dimensional GMM (M mixtures, $\Lambda = \{\omega_m, \mu_m, \sigma_m^2\}, m = 1, \ldots, M$) using the resulting scores. The prior parameters are then calculated as [28]

$$\mu = \sum_{m=1}^{M} \omega_m \mu_m \tag{18}$$

$$\alpha = \frac{1}{\sigma^2} = \frac{1}{\sum_{m=1}^{M} \omega_m \sigma_m^2} \tag{19}$$

$$r = \frac{1}{\alpha \sum_{m=1}^{M} \omega_m (\mu_m - \mu)^2} \tag{20}$$

$$\beta = 1. \tag{21}$$

The above settings ensure that in the absence of any observed samples the estimated mean and variance are a weighted average of the component means and variances in the constructed GMM.

In our experiments, we found that the development utterances resulted in a very uneven selection of the top scoring in-set speaker. So, in addition to the above technique, two other schemes were employed to estimate the mean and variance parameters for the in-set speakers:

- In case the number of samples available for MAP estimation falls below a certain number (100 in our experiments), use the Top-norm estimate for the mean and variance.
- Determine only the mean estimate via MAP; choose the prior mean as the Top-norm (or Z-norm) mean; set the variance the same as the variance of the Top-norm or Z-norm scheme and empirically determine a relevance factor (r in the above equations) to control the balance of adaptation.

V. Experiments

A. Experimental Setup

1) Timit: A set of 60 male speakers were randomly selected as the in-set/out-of-set speaker space. These 60 speakers serve both as in-set speakers and out-of-set speakers (impostors) 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 in-set/out-of-set speaker 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 train/test data overlap and initial/trailing silence removed. The training data was limited to approximately 5 s of speech, while test data was created for 2, 4, 6, and 8 s of speech. The remaining 378 male speakers, each having about 30 s of data, were used as development data.

2) CU-Move: The CU-Move speech corpora consists of speech and noise data collected from within vehicles under a variety of driving conditions, in order to facilitate design of in-vehicle interactive systems for route planning and navigation [29]–[31]. For this study, we have selected part of the corpus

[1]RAD/MAX/UCN derive their names form the specific techniques used during score normalization, please refer to [24] or [5] for a detailed description.

containing phonetically balanced TIMIT sentences that were read in a moving vehicle. The setup of the in-set and out-of-set speakers and the training data duration is identical to the TIMIT setup above. Test data durations consisted of 2, 3, 4, and 6 s of speech. One-hundred and fifty-six male speakers were used as development data (a total of 4.8 hours of speech).

3) NIST-SRE: The NIST-SRE database is made up of conversational telephony speech. With attendant noise and channel variations, this represents a very challenging environment for speaker recognition systems. Since a speaker verification system performs considerably better than an in-set system [3][2] and the cohort-based speaker modeling scheme (Section III) can be applied to the speaker verification case as well, speaker verification experiments were carried out using the 10sec train/10sec test NIST-SRE 2005 data set for male speakers. SRE 2004 data from male speakers was used for speaker modeling.

For experiments on score normalization, an in-set/out-of-set system was constructed using 40 speakers from the one-conversation train/test condition (slightly more than 2 min of speech for training and testing). A round-robin test procedure using four groups of 10in/30out, 20in/20out and 30in/10out speakers was employed. A total of 1000 SRE 2005 (male) single conversation test files were used as development data for estimation of the score normalization parameters. SRE 2004 data from male speakers was used for UBM construction.

B. Front-End Processing

The speech analysis frame rate was set to 30 ms for TIMIT and 20 ms for CU-Move and NIST-SRE datasets, with a 10-ms skip rate and the speech utterance was preemphasized with the filter $(1-0.95z^{-1})$. Nineteen-dimensional Mel-frequency cepstral coefficients (MFCCs) were extracted and used for statistical modeling. Silence and low-energy speech parts were removed using an energy-based frame selection technique.

C. Speaker Modeling

For all three corpora, based on the performance on the evaluation set, the size of the GMM is chosen such that the baseline algorithm achieves the highest performance (among competing GMM sizes). Similarly, after experimenting with a number of cohort set sizes (N_{cohort}), the top performing size was chosen for all in-set speakers.[3] The speakers are modeled as follows.

- *TIMIT* ($N_{\text{cohort}} = 10$): A UBM containing 32 Gaussian components is constructed by randomly selecting 60 speakers (all male) from the development set. The remaining 318 (378–60) male speakers are used for selecting the cohort sets for the in-set speakers. In experiments using upto 256 Gaussians for the GMMs and all 378 speakers for baseline UBM construction, the baseline performance was at about the same level. MAP adaptation is performed with a fixed relevance factor of 16 for all speaker models. Fig. 2 shows the influence of N_{cohort} on performance for a test data duration of 4 s.

[2]We found that an in-set system constructed using the 10sec train, 10sec test Speaker Recognition Evaluation (SRE) data gave a very low baseline score, with performance approaching that of a random detector.

[3]This is also the procedure followed while selecting other hyperparameters, e.g., M, r in (18)–(21).

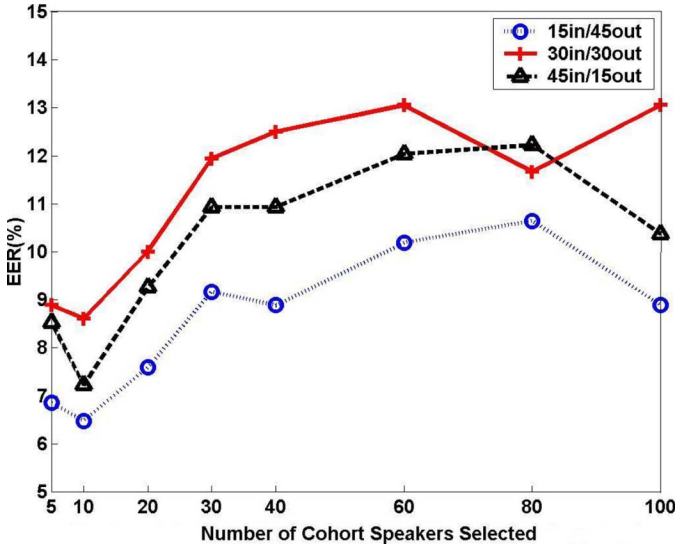

Fig. 2. Equal error rate (%EER) results for the proposed algorithm on TIMIT, for different cohort set sizes (N_{cohort}) and test data duration of 4 s.

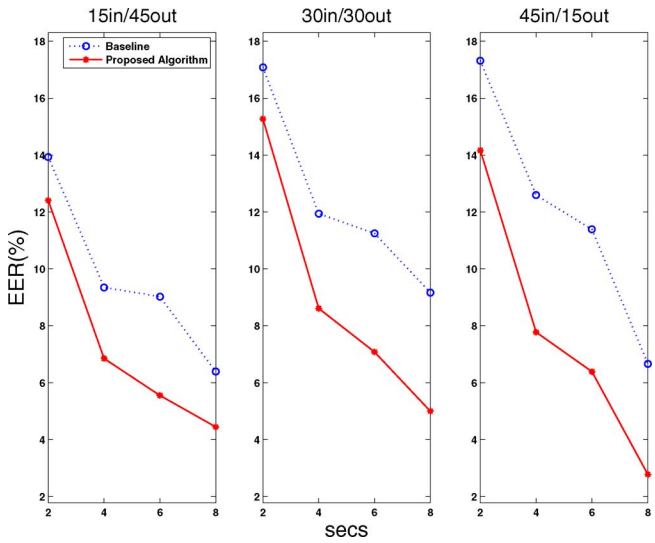

Fig. 3. Performance [in terms of EER(%)] of baseline and proposed algorithm on TIMIT, using in-set/out-of-set speaker sizes of 15/45, 30/30, and 45/15.

- *CU-Move* ($N_{\text{cohort}} = 8$): A UBM containing 64 Gaussian components is constructed using all the 156 development speakers. The baseline system uses MAP estimation with a fixed relevance factor of 16. Cohort sets are constructed using the 156 speakers, with mixture specific relevance factors [(13)].
- *NIST-SRE* ($N_{\text{cohort}} = 20$): A UBM containing 64 Gaussian components is constructed using SRE 2004 training data. MAP adaptation is performed with a fixed relevance factor of 16 for all speaker models. For the score normalization experiments, the size of the UBM is 256.

The steps presented in Section III-B were performed to obtain the speaker models using the proposed algorithm.

D. Evaluations

Fig. 3 shows the resulting EER for the baseline and proposed systems for three different in-set/out-of-set configurations of

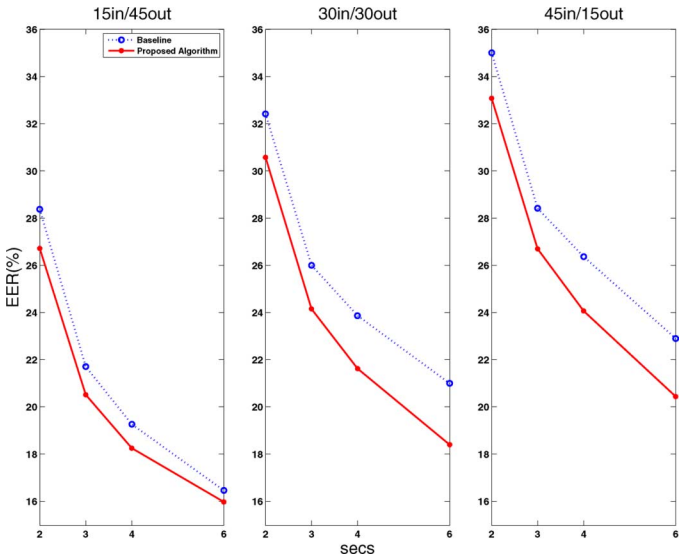

Fig. 4. Performance [in terms of EER(%)] of baseline and proposed algorithm on CU-Move, using in-set/out-of-set speaker sizes of 15/45, 30/30, and 45/15.

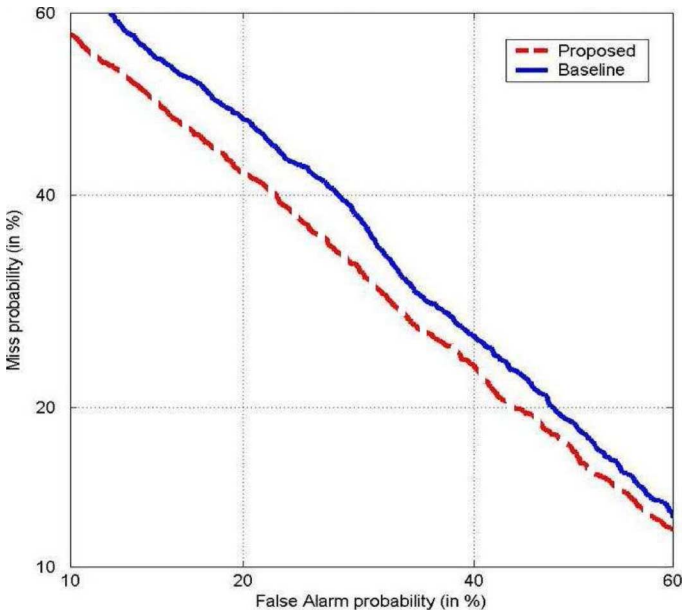

Fig. 5. DET curve for NIST-SRE 2005 10 s/10 s condition (male speakers only) showing performance of baseline and proposed algorithm.

15in/45out, 30in/30out and 45in/15out, for varying duration of test data. Fig. 4 shows the corresponding plot for the CU-Move database from in-vehicle recordings.

From Figs. 3 and 4, consistent and in some cases, especially for the 6- and 8-s test conditions, significant improvement in performance is observed. As noted in our previous in-set/out-of-set studies [5]–[7], [14], results for test sets of size 2 s perform measurably different than 4–8 s test sets. In general, higher (relative) improvements always occur for longer duration test sets.

Fig. 5 shows the detection error tradeoff (DET) plot for a speaker verification system using the baseline and proposed schemes with train and test data durations of 10 s each. A slight but consistent improvement in performance is seen over all operating points. EER improves from 32.686% to 30.658%.

TABLE II
PERCENTAGE EER VALUES FOR THE DIFFERENT IN-SET/OUT-OF-SET SIZES CONSIDERED. SPEAKERS ARE SELECTED FROM THE 1CONV/1CONV SRE-2006 DATABASE. ALL SYSTEMS USE THE UBM-BASED LRT FOR INITIAL SCORE NORMALIZATION

System EER's	10in/30out	20in/20out	30in/10out
Baseline	27.961	31.830	33.834
Z-norm	26.285	28.771	32.298
Top-norm	24.787	27.724	31.311
Proposed algorithm	27.858	31.245	33.114

Table II shows the EER obtained for the in-set/out-of-set speaker configurations of 10in/30out, 20in/20out, and 30in/10out, using different score normalization schemes. The baseline scores are obtained using the UBM-based LRT [(4)]. The Z-norm parameters are computed using (15), and these are used to normalize the baseline scores. Top-norm parameters are computed by using the development utterances contributing to the top 10% of the scores. For the proposed algorithm (17) is evaluated, (18)–(21) are computed with a GMM size of 8. For the situation where the number of samples available for adaptation is below 100, the corresponding top-norm estimates are used.

VI. DISCUSSION

It is clear that with 5 s of training data, acoustic holes in the speaker production space will be present. This has been observed in earlier studies [5]–[7]. The proposed algorithm provides measurable improvements across all corpora and test conditions. Significant improvement is seen for the 6–8 s test sizes for the TIMIT corpus (Fig. 3, with relative improvement in EER in the range of 30.43–58.33%). Since this database is noise-free and speaker data is collected in a single recording session, this represents the best possible operating environment for speaker recognition algorithms. The improvements seen on noisier corpora are not as high. In the noisy in-vehicle CU-Move corpora, the highest relative improvement is 12.38% for the 30in/30 out 6-s test set case (Fig. 4). Since for both the TIMIT and CU-Move databases the order of improvement for the 2-s case is quite different from the 4–8 case, this suggests that for a test size of 2 s, the classifier structure should be different than for the 4–8 test sizes. For the speaker verification experiment on the NIST-SRE corpora, consistent improvement is seen (Fig. 5). Since we have performed no channel compensation, it is likely that development speakers whose channel, rather than acoustic, characteristics match will be selected as cohorts for an enrolled speaker, hence the impact on performance is not large. A limitation of the proposed cohort-based speaker model scheme is that if very few speakers in the development speaker set are similar to a given in-set speaker, then the cohort-based speaker model for that in-set speaker will be a very poor representation for that speaker and since more out-of-set speakers in this in-set speaker's ML decoding region will get accepted, the error rate for that speaker will be higher compared to the baseline.

Some directions for future work are as follows.

- Currently, the size of the cohort set is fixed at the same number for all in-set speakers. Performance should be better if this size were chosen separately for each in-set speaker depending on, for example, the cross-verification

scores. This would ensure a more consistent measure of similarity for the cohort set selected for the in-set speakers.
- The experiments did not use any environment/channel compensation schemes, this could lead to an erroneous selection of cohorts for an enrolled speaker. By using existing compensation schemes, more performance improvement can be expected. Also, this would allow sharing of speaker data across databases.
- The performance of the method using various in-set specific score normalization criteria [5], [19] could also be investigated.

For the score normalization experiments, the main challenge we faced was a lack of diversity in the development data. Using 1000 development utterances for an in-set size of 30 speakers, many in-set speakers had less than ten samples for computing their score normalization parameters. Though the MAP-based scheme attempted to alleviate this, the lack of diversity in the development utterances is a major factor in limiting performance. Future work should concentrate on addressing this issue.

VII. Conclusion

In this paper, we have studied the problem of text-independent in-set/out-of-set speaker recognition, with extremely short-duration enrollment and test data sizes. We proposed an algorithm that uses an in-set speaker's cohort set to make up for the sparse (e.g., 5 s per speaker) enrollment data. Investigations on a clean speech database show significant improvement for the proposed method over a GMM–UBM baseline. Evaluations using the noisy in-vehicle CU-Move corpus and the NIST-SRE database show a consistent improvement in performance. We also formulated an in-set/out-of-set system specific score normalization scheme and compared it with existing score normalization schemes. Future work will primarily focus on choice and composition of the cohort-set and the use of score normalization techniques (other than world-model based). Additionally, we will look to mitigate the impact of lack of diversity of development utterances for the proposed score normalization scheme.

Acknowledgment

The authors would like to thank J.-W. Suh of CRSS, University of Texas at Dallas, for extraction and setup of the CU-Move corpus. They would also like to thank the anonymous reviewers whose comments significantly improved the quality of the manuscript.

References

[1] J. H. L. Hansen, R. Huang, B. Zhou, M. Seadle, J. R. Deller, A. R. Gurijala, M. Kurimo, and P. Angkititrakul, "Speechfind: Advances in spoken document retrieval for a national gallery of the spoken word," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 712–730, Sep. 2005.

[2] J. Campbell, "Speaker recognition: A tutorial," *Proc. IEEE*, vol. 85, no. 9, pp. 290–294, Sep. 1997.

[3] E. Singer and D. Reynolds, "Analysis of multitarget detection for speaker and language recognition," in *Proc. Odyssey 2004 Speaker Lang. Recognition Workshop*, 2004, pp. 301–308.

[4] P. Philips, P. Grother, R. Micheals, D. Blackburn, E. Tabassi, and J. Bone, "FRVT 2002: Evaluation Report." Tech. Rep. 2003 [Online]. Available: http://frvt.org/FRVT2002/documents.htm

[5] P. Angkititrakul and J. H. L. Hansen, "Discriminative in-set/out-of-set speaker recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 498–508, Feb. 2007.

[6] P. Angkititrakul, J. H. L. Hansen, and S. Bagahaii, "Cluster-dependent modeling and confidence measure processing for in-set/out-of-set speaker identification," in *Proc. Odyssey 2004 Speaker Lang. Recognition Workshop*, 2004, pp. 2385–2388.

[7] P. Angkititrakul and J. H. L. Hansen, "Identifying in-set and out-of-set speakers using neighborhood information," in *Proc. ICASSP'04*, 2004, pp. 393–396.

[8] A. Higgins, L. Bahler, and J. O. Porter, "Speaker verification using randomized phrase prompting," *Digital Signal Process.*, vol. 1, pp. 89–106, 1991.

[9] A. E. Rosenberg, J. DeLong, C.-H. Lee, B.-H. Juang, and F. K. Soong, "The use of cohort normalized scores for speaker verification," in *Proc. ICSLP'92*, 1992, pp. 599–602.

[10] D. A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," in *Eurospeech'97*, 1997, pp. 1379–1382.

[11] W. Liu, T. Isobe, and N. Mukawa, "On optimum normalization method used for speaker verification," in *Proc. ICSLP'98*, 1998, paper 1045.

[12] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," in *Digital Signal Process.*, 2000, vol. 10, pp. 42–54.

[13] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process.*, vol. 10, pp. 19–41, 2000.

[14] V. Prakash and J. H. L. Hansen, "A cohort – UBM approach to mitigate data sparseness for in-set/out-of-set speaker recognition," in *Proc. Interspeech'06*, 2006, paper 1847.

[15] Y. Mami and D. Charlet, "Speaker modeling from selected neighbors applied to speaker recognition," *Eurospeech'03*, pp. 2629–2632, 2003.

[16] D. Charlet, "Neighborhood-adapted GMM for speaker recognition," in *Proc. Odyssey 2004 Speaker Lang. Recognition Workshop*, 2004, pp. 227–230.

[17] M. Ben, R. Blouet, and F. Bimbot, "A Monte Carlo method for score normalization in automatic speaker verification systems using Kullback–Leibler distances," in *Proc. ICASSP'02*, 2002, pp. 689–692.

[18] V. Prakash and J. H. L. Hansen, "Score distribution scaling for speaker recognition," in *Proc. Interspeech*, Antwerp, Belgium, Aug. 2007.

[19] P. Sivakumaran, J. Fortuna, and A. M. Ariyaeeinia, "Score normalisation applied to open-set, text-independent speaker identification," in *Eurospeech'03*, 2003, pp. 2669–2672.

[20] J. Fortuna, P. Sivakumaran, A. M. Ariyaeeinia, and A. Malegaonkar, "Open-set speaker identification using adapted Gaussian mixture models," in *Proc. Interspeech'05*, 2005, pp. 1997–2000.

[21] J. Fortuna, P. Sivakumaran, A. M. Ariyaeeinia, and A. Malegaonkar, "Relative effectiveness of score normalisation methods in open-set speaker identification," in *Proc. Odyssey 2004 Speaker Lang. Recognition Workshop*, 2004, pp. 369–376.

[22] Y. Zigel and M. Wasserblat, "How to deal with multiple-targets in speaker identification systems?," in *Proc. IEEE Odyssey 2006 Speaker Lang. Recognition Workshop*, 2006, pp. 1–7.

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

[24] J. Schurmann, *Pattern Classification: A Unified View of Neural and Statistical Approaches*. New York: Wiley, 1996.

[25] H. Jiang and L. Deng, "A Bayesian approach to the verification problem: Applications to speaker verification," *IEEE Trans. Speech Audio Process.*, vol. 9, pp. 874–884, Nov. 2001.

[26] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. A. Reynolds, "Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation," in *Proc. ICSLP'98*, 1998, paper 0608.

[27] M. H. DeGroot, *Optimal Statistical Decisions*. New York: McGraw-Hill, 1970.

[28] C.-H. Lee, C.-H. Lin, and B.-H. Juang, "A study on speaker adaptation of the parameters of continuous density hidden Markov models," *IEEE Trans. Signal Process.*, vol. 39, no. 4, pp. 806–814, Apr. 1991.

[29] J. H. L. Hansen, P. Angkititrakul, J. Plucienkowski, S. Gallant, U. Yapanel, B. Pellom, W. Ward, and R. Cole, "CU-move : Analysis and corpus development for interactive in-vehicle speech systems," in *Proc. Eurospeech'01*, 2001, pp. 2023–2026.

[30] J. H. L. Hansen, J. Plucienkowski, S. Gallant, B. Pellom, and W. Ward, "CU-move: Robust speech processing for in-vehicle speech systems," in *Proc. ICSLP'00*, 2000, pp. 524–527.

[31] J. H. L. Hansen, X. X. Zhang, M. Akbacak, U. H. Yapanel, B. Pellom, W. Ward, and P. Angkititrakul, "CU-MOVE: Advanced In-Vehicle Speech Systems for Route Navigation," in *DSP for In-Vehicle and Mobile Systems*, H. Abut, J. H. L. Hansen, and K. Takeda, Eds. New York: Springer, 2004.

Vinod Prakash (S'04) received the B.E. degree in electronics and communications engineering from Karnataka Regional Engineering College, Surathkal, India, in 2001 and the M.S. degree in electrical engineering from the University of Colorado at Boulder, Boulder, CO, in 2007.

He is currently a Software Design Engineer with Microsoft Corporation, Redmond, WA, working on the topic of acoustic echo cancellation. From 2001 to 2004, he was a Senior Engineer at Ittiam Systems Pvt., Ltd., Bangalore, India, and worked in the area of audio compression technologies. From summer 2005 to fall 2006, he was a Research Intern at the Center for Robust Speech Systems (CRSS), University of Texas at Dallas, Richardson, working on the topic of speaker recognition. His research interests include automatic speech/speaker recognition, statistical signal processing and pattern recognition.

John H. L. Hansen (S'81–M'82–SM'93–F'07) received the B.S.E.E. degree from the College of Engineering, Rutgers University, New Brunswick, NJ, in 1982 and the M.S. and Ph.D. degrees in electrical engineering from the Georgia Institute of Technology, Atlanta, in 1983 and 1988, 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 a Department Chairman of Electrical Engineering and holds the Distinguished University Chair in Telecommunications Engineering. He also holds a joint appointment as Professor 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, at the University of Colorado at Boulder, 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 39 (18 Ph.D., 21 M.S.) thesis candidates. He is author/coauthor of 247 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), *Advances for In-Vehicle and Mobile Systems: Challenges for International Standards* (Springer, 2007), and lead author of the report "The Impact of Speech Under 'Stress' on Military Speech Technology," (NATO RTO-TR-10, 2000).

Prof. Hansen served as an IEEE Signal Processing Society Distinguished Lecturer for 2005/2006, is a member of the IEEE Signal Processing Society Speech Technical Committee and Educational Technical Committee, and has served as Technical Advisor to the U.S. Delegate for NATO (IST/TG-01). He was an Associate Editor for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (1992–99), an Associate Editor for IEEE SIGNAL PROCESSING LETTERS (1998–2000), and an 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 TRANSACTION ON 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 (2004–2010). He was a 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 IEEE ICASSP-2010, Dallas, TX.