Acoustic hole filling for sparse enrollment data using a cohort universal corpus for speaker recognition

Jun-Won Suh and John H. L. Hansen[a)]

Center for Robust Speech Systems (CRSS), The University of Texas at Dallas, 800 W. Campbell Road, Richardson, Texas 75080

(Received 4 December 2010; revised 1 December 2011; accepted 5 December 2011)

In this study, the problem of sparse enrollment data for in-set versus out-of-set speaker recognition is addressed. The challenge here is that both the training speaker data (5 s) and test material (2 ~ 6 s) is of limited test duration. The limited enrollment data result in a sparse acoustic model space for the desired speaker model. The focus of this study is on filling these acoustic holes by harvesting neighbor speaker information to leverage overall system performance. Acoustically similar speakers are selected from a separate available corpus via three different methods for speaker similarity measurement. The selected data from these similar acoustic speakers are exploited to fill the lack of phone coverage caused by the original sparse enrollment data. The proposed speaker modeling process mimics the naturally distributed acoustic space for conversational speech. The Gaussian mixture model (GMM) tagging process allows simulated natural conversation speech to be included for in-set speaker modeling, which maintains the original system requirement of text independent speaker recognition. A human listener evaluation is also performed to compare machine versus human speaker recognition performance, with machine performance of 95% compared to 72.2% accuracy for human in-set/out-of-set performance. Results show that for extreme sparse train/reference audio streams, human speaker recognition is not nearly as reliable as machine based speaker recognition. The proposed acoustic hole filling solution (MRNC) produces an averaging 7.42% relative improvement over a GMM-Cohort UBM baseline and a 19% relative improvement over the Eigenvoice baseline using the FISHER corpus. © *2012 Acoustical Society of America*. [DOI: 10.1121/1.3672707]

PACS number(s): 43.72.Fx [CYE] Pages: 1515–1528

I. INTRODUCTION

A major challenge for effective speaker recognition occurs when the enrollment data are very sparse. Limited enrollment data cause a severe lack of phone coverage for those phonemes that are rarely seen in the training data but seen in the test sequence. This phenomenon will be referred to as "acoustic holes" in the speaker acoustic model space. The acoustic holes from the sparse enrollment speaker's data can be filled by borrowing phoneme data from development speakers (Prakash and Hansen, 2007) that are acoustically similar, so it is important to select acoustically close development speakers for each enrolled speaker. A novel speaker similarity measurement plays an important role in recruiting similar cohort speakers. In this study, three different speaker similarity measures are employed at: the feature level, feature and statistical model level, and statistical model level.

Another potential direction for improving performance of speaker recognition constrained with sparse enrollment data is to employ a statistical model to characterize the specific speaker traits. Many researchers have studied statistical algorithms for building robust speaker models using sparse enrollment data (Prakash and Hansen, 2007; Angkititrakul and Hansen, 2007; Mak *et al.*, 2006) based on vector quantization (VQ) (Soong *et al.*, 1985), Eigenvoice (Kuhn *et al.*, 2000; Kenny *et al.*, 2007), and support vector machine (SVM) (Burges, 1998). It is noted that the Gaussian Mixture Model (GMM) for speaker modeling has become a well-known algorithm for building robust text-independent speaker groups, known as the universal background model (UBM) (Reynolds *et al.*, 2000). The GMM-UBM algorithm represents speaker models for many scenarios based on an accurate speaker space assuming sufficient training data. However, effective speaker adaptation is necessary for the case of sparse enrollment data. In a previous study, structural speaker adaptation was used for compensating the sparse enrollment and claimant speaker's data (Matrouf *et al.*, 2003). In a GMM, the enrolled speaker dependent model is formulated by adapting a Gaussian mixture from a UBM with enrollment data via maximum a posteriori (MAP) adaptation (Reynolds *et al.*, 2000; Gish and Schmidt, 1994). In this study, two effective speaker adaptation methods are presented that incorporate the parallel speaker selection process to provide effective data coverage, therefore accurately representing the given speaker acoustic space.

A number of scenarios are possible for the general problem of speaker recognition. One configuration that has recently emerged is *in-set/out-of-set* speaker recognition. Here, the goal is simply to determine if an input speaker is a member of an in-set group or if the speaker is rejected (Angkititrakul and Hansen, 2007). The in-set/out-of-set speaker recognition scenario draws from two speaker recognition applications; (i) closed-set speaker identification

[a)] Author to whom correspondence should be addressed. Electronic mail: john.hansen@utdallas.edu

(Gish and Schmidt, 1994) followed by (ii) open-set speaker verification (Reynolds *et al.*, 2000; Li *et al.*, 2003). Closed-set speaker identification identifies speech from the claimant speaker among the set of enrolled speaker models, and open-set speaker verification produces a yes/no determination for the claimant speaker's speech. The combination of these two applications enables in-set/out-of-set recognition to increase computation speed and in-set detection accuracy with sparse data. Speaker verification uses only the claimant speaker's data and model to produce a binary decision, and sparse data can produce an unreliable decision caused by spare phone coverage. The in-set/out-of-set recognition scenario increases the size of the enrollment speaker set so that the system exploits a set of enrolled speaker models instead of only one speaker model. The set of speaker models is grouped as in-set speakers, and the claimant speaker is identified within the in-set speakers using closed-set speaker identification. The proposed acoustic-hole filling algorithm is evaluated in the context of an *in-set/out-of-set* scenario to measure performance. Because in-set/out-of-set speaker recognition identifies one subject as belonging to a group of speakers, it can be applied in spoken document retrieval (Hansen *et al.*, 2005), for audio library or broadcast news programs, or in security applications to capture audio for specific groups of individuals.

Human speaker recognition using sparse data is considered for in-set/out-of-set versus open-set speaker identification (OSI) (Gish and Schmidt, 1994) to study the benefit of in-set/out-of-set knowledge for not only machine, but also human listening performance. It is difficult to determine which performs better for the task of speaker recognition in human versus machine because there are many factors that impact performance, such as the data size of enrolled speaker, speaker familiarity to the listeners, personal listener characteristics (e.g., familiarity with accents, emotions, etc.), and speech content of the data (Furui, 1997). It is noted that human performance is better at speaker recognition than machine using speech degraded by background noise or alternative handsets (Schmidt-Nielsen and Crystal, 2000), suggesting that human listening ability effectively interpolates missing information when data is limited. Because the focus of this study is on sparse data, it would be useful to explore human listening ability in overcoming acoustic holes by comparing human versus machine performance.

A. Sparse data analysis in speaker recognition

Limited phone coverage caused by sparse data from the enrolled speaker makes speaker recognition vulnerable to unseen phones from a claimant speaker (i.e., a "correct" speaker would be rejected due to phoneme space gaps between train and test phonemes; a speaker who should be "rejected" is more likely to be accepted if that speaker produces a high overlap between their two phoneme spaces). Developing effective ways to increase the available enrollment data should in theory improve speaker recognition performance because of an improvement in the balance of phone coverage for the enrolled speaker model. Limited phone coverage can be enriched by borrowing data from acoustically close speakers assuming those speakers would be unlikely input claimants, and an analysis of the given phone coverage is required to determine effective data requirements for filling the acoustic holes (i.e., because these acoustic holes are being filled by outside but hopefully acoustically similar speakers, it is important to use as little outside data as possible to fill the holes but not adversely migrate the given speaker model towards the parallel hole filling speakers.). The proposed method will therefore fill acoustic holes by borrowing data from acoustically close speakers after phone coverage analysis.

The distribution of phone occurrence in conversational speech is different for each phoneme. Figure 1 shows results from a study on how the distribution of phonemes decreases with a reduction in the conversational speech time from 10 to 1 min. The phone occurrence distribution for an average 10 min speaker conversation from 26 speakers is reported in Mines *et al.* (1978), and phone duration levels for 1 min of speech are reduced assuming a constant overall time ratio in order to illustrate the reduction of phone occurrence. The dramatic reduction of data will create "complete" acoustic holes for missing phones and "half-filled" acoustic holes for those phones which are less frequently seen. It is necessary to know the amount of phone coverage needed to fill complete and half-filled acoustic holes from acoustically close speakers, and an ideal phone coverage phone table can show this information. In this study, each specific phone type is replaced by a general Gaussian mixture, and a mixture tagging table is employed instead of the exact phone coverage table (e.g., this is done because of the text independent requirement).

It is possible to employ a phoneme recognizer to classify the input phoneme sequence within the available utterance, but such a recognizer will have its own phone recognizer error rate, typically 25% ~ 35% even in noise-free conditions (Lee and Hon, 1989; Lamel and Gauvain, 1993). Using a GMM mixture tagging strategy allows for phone-like

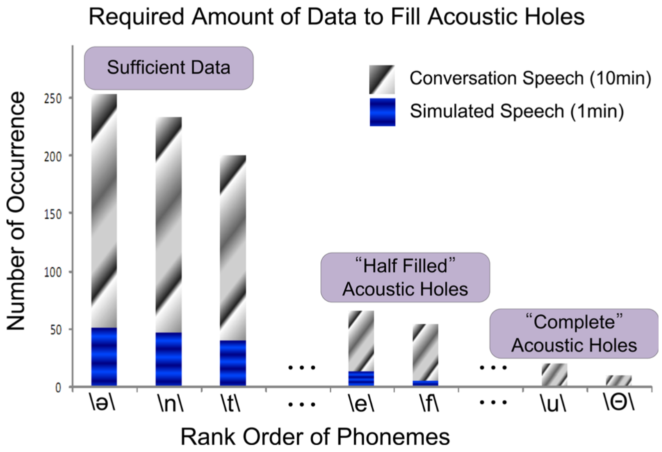

FIG. 1. (Color online) This plot illustrates the rank order of frequency of occurrence of each phoneme count of 4000 phoneme tokens (≈10 min) from conversational speech (Mines *et al.*, 1978). The plot also shows the impact on "Simulated Speech (1 min)" phoneme count. "Sufficient *data*" indicates enough data to manipulate acoustic space for corresponding phoneme. "Complete Acoustic hole" indicates required phone data to form a complete acoustic space for balanced speaker model.

labeling without the specific requirement of knowing which GMM mixtures correspond to each phoneme. It should be noted, however, that the goal here is to balance the phoneme coverage for short enrollment data sets to that comparable for larger training sets. We recognize that the resulting balancing process uses the exact optimal coverage for the phoneme space. However, the advantage of using an unsupervised GMM mixture tagger in place of a phoneme recognizer, which will have a high error rate, is the key benefit being leveraged. Here, a GMM is employed to classify the speaker information that contains phonemes and intrinsic/extrinsic information (Torres-Carrasquillo *et al.*, 2002; Scheffer and Bonastre, 2006). It is assumed that the GMM characterizes the speaker independent acoustic information because this model is trained with approximately 400 speaker utterances. Each speaker dependent trait will merge with closely related characteristic classes in a large number of utterances, and it is expected that the total mixture number of the GMM will represent the phoneme acoustic space for the set of independent speakers. A mixture tagged histogram is constructed for counting the number of mixture occurrences for the development feature data to imitate the proper frequency of phone occurrence (e.g., a rank order illustration is shown conceptually in Fig. 1). This mixture occurrence counting table is referred to as the "balanced mixture tagged histogram." A cohort model for each enrolled speaker is formed with the same phone dependent mixture index data as reflected in the balanced mixture tagged histogram. Figure 2 shows the system goal for speaker modeling to build a balanced acoustic space. To illustrate the incredible difficulty of the proposed task (e.g., sufficient training with 600 s vs. limited data of 5 s), an example is considered based on images and human visual recognition. Assume an image of a well known individual can be partitioned into 600 image blocks [Fig. 3 (a)]. With a training data size of 600 s, the acoustic model can represent the complete "acoustic image" [Fig. 3(a)]. When the test size is 2–6 s, the number of sample blocks will be 2–6 [see Fig. 3(c) and 3(d)]. While this is limited, reasonable performance can be achieved because of the accurate trained base model. However, with a speech training size of 5 s [represented as image Fig. 3(b)], the image is almost random.

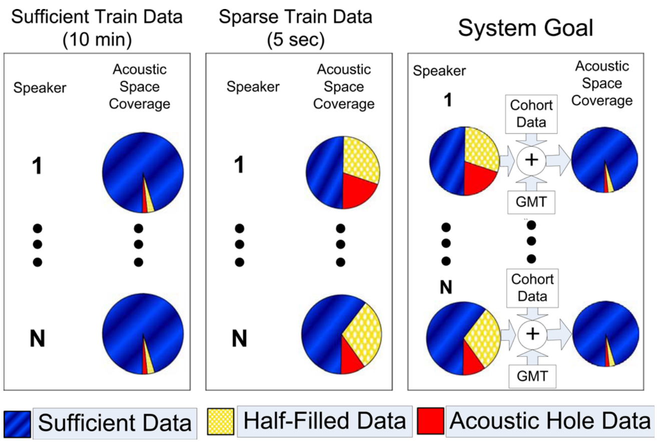

FIG. 2. (Color online) This plot illustrates the goal of this study. The acoustic holes (half-filled/complete holes) are appeared by sparse data, and the acoustic holes are filled by borrowing acoustically similar speakers. The GMT from Sec. III is used to tag the acoustic information, and the speaker model covers the acoustically balanced space.

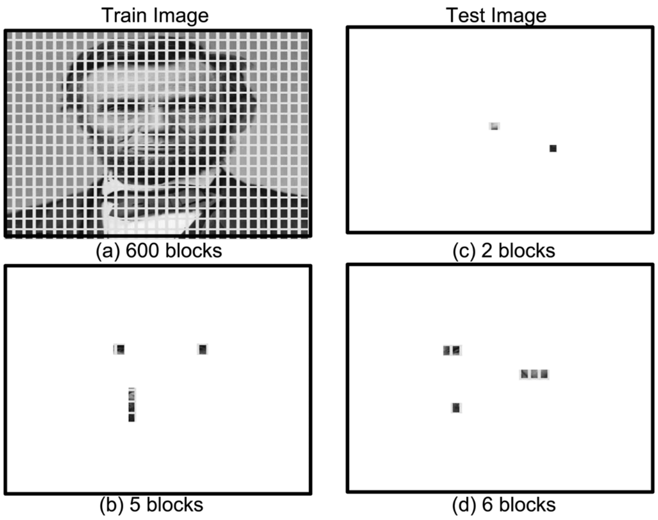

FIG. 3. President Lincoln image to illustrate the problem of speaker recognition with spare data.

When the test size is 2–6 s, the number of test blocks becomes 2–6 (Fig. 3), illustrating how truly challenging the pattern recognition task will be. In the following sections, an elegant data selection scheme for supporting sparse enrollment data is proposed.

The remainder of this paper is organized as follows. The baseline in-set/out-of-set speaker recognition system is introduced in Sec. II. The phoneme related mixture classification GMM scheme is described in Sec. III. Section IV focuses on introducing three different speaker similarity measurement selection methods. The proposed strategy for speaker modeling is presented in Sec. V, which includes a speaker adaptation scheme (Sec. V A), and a smart data selection method for cohort modeling in Sec. V B. Section VI describes the evaluation and results of the proposed algorithms employing FISHER and CU-MOVE corpora as well as a separate human listener evaluation. Section VII includes further analysis on the experimental results and considers some general issues on the use of multi-channel corpora. Finally, Sec. VIII summarizes the contributions, draws conclusions, and suggests directions for future work.

II. BASELINE SPEAKER RECOGNITION SYSTEM

A. In/out-of-set speaker recognition

The basic concepts of in-set/out-of-set speaker recognition are described in this section, and more complete discussion can be found in Angkititrakul and Hansen (2007). It can be applied to audio search or speaker diarization by selecting target speaker as in-set group for your understanding of in-set/out-of-set speaker recognition in application aspect. 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 $\mathbf{X}_n = \{\mathrm{x}_{n1}, \mathrm{x}_{n2}, \ldots, \mathrm{x}_{nT_n}\}$ 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 the following two steps to be performed. In the first stage, called *(closed-set) speaker identification* or *speaker classification,* X is first classified into one of the most likely in-set speakers Λ^* as,

$$\Lambda^* = \underset{1 \leq n \leq N}{\operatorname{argmax}} \; p(\mathrm{X}|\Lambda_n). \tag{1}$$

In the second stage, called *speaker verification* or *outlier verification,* it verifies 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 $\mathbf{H}'$, which represents the hypothesis where $\mathbf{X}$ is actually *not* the speaker model Λ^*. If the probabilities of the null and alternative hypotheses are assumed known, then according to the Neyman–Pearson Lemma, the conventional likelihood ratio test (LRT) is optimal (Huber, 1965) (in terms of correct detection for a specific false alarm),

$$\frac{p(\mathrm{X}|\Lambda^*)}{p(\mathrm{X}|\Lambda_0)} \begin{cases} \geq \gamma : \text{accept H} \\ < \gamma : \text{reject H (accept H}'), \end{cases} \tag{2}$$

where γ is a pre-defined threshold, Λ_0 is a speaker independent model (e.g., UBM or cohort-speaker models), and $p(\cdot|\cdot)$ is the likelihood given each speaker model Λ. The competitive or antispeaker model score, $p(\mathbf{X}/\Lambda_0)$, can be computed using the UBM (Reynolds *et al.,* 2000), unconstrained cohort model (Sivakumaran *et al.,* 2003), and MAX rule (Kressel and Schurmann, 1997),

$$\text{UCN rule} : p(\mathrm{X}|\Lambda_0) = \frac{1}{N}\sum_{n=n-1}^{N} p(\mathrm{X}|\Lambda_n) \tag{3}$$

$$\text{MAX rule} : p(\mathrm{X}|\Lambda_0) = \max_{1 \leq n \leq N, \Lambda_n \neq \Lambda^*} p(\mathrm{X}|\Lambda_n). \tag{4}$$

It is possible to predict system performance based on different in-set group sizes by employing a fixed anti-speaker model such as a UBM. In a previous study (Angkititrakul and Hansen, 2007), it was shown that as the in-set group size becomes larger, the out-of-set score distribution will move toward the in-set score distribution (e.g., becomes more confusable). If it is assumed that the true claimant produces the highest probability for the claimant speaker in the closed speaker identification stage, then the in-set/out-of-set system will have achieved the desired task goal. A smaller sized in-set group will perform better than a larger size in-set group. This predictable result is based on the observation that the true claimant speaker model produces the best probability score. For example, if the true claimant enters the system, there are a fixed number of in-set probability scores from the numerator of Eq. (2) and a fixed UBM model score, and therefore the LRT does not have an effect on the in-set group size. Even if one member of the in-set group produces the highest probability (even if it is not the same in-set speaker), the system performance will still be a positive effect. When a false claimant enters the system, an evaluation with as few as 15 in-set speakers has less speaker models to compare with the false claimant. When the in-set size is increased to 45, in-set speaker models are generally closer for the false claimant because there are more speaker models than the 15 in-set to select from group. The numerator part from Eq. (2) has a higher probability when there are 45 in versus 15 in, and therefore the out-of-set score distribution moves toward the in-set score distribution. While this example is a simplified version, the net idea is that the in-set speakers will have a higher probability from Eq. (1), and therefore the outlier model plays an important role in shifting the score distribution. It should be noted that the score distribution can also be normalized by Eq. (3) or (4).

B. Baseline GMM-UBM system

Here a baseline system is needed for comparison to illustrate the proposed acoustic hole filling methods. The baseline system used in Angkititrakul and Hansen (2007) is also employed here with a brief overview presented in this section. We employ a GMM as the UBM with MAP speaker adaptation for our text-independent speaker recognition baseline (Reynolds *et al.,* 2000). A speaker independent model, or UBM, is trained from non-target speakers using the expectation maximization (EM) algorithm. The speaker model is represented by M Gaussian components trained from a sequence of D dimensional observation vectors x_t. The GMM is denoted as $\Lambda_n = (\omega_{nm}, \mu_{nm}, \Sigma_{nm})$, for $m = 1, \ldots, M$ and $n = 1, \ldots, N$, where M is the number of mixtures, and N the number of speakers. ω_{nm} is the mixture weight of the mth component unimodal Gaussian density $\mathcal{N}_{nm}(\mathrm{x_t})$, where each is parameterized by a mean vector μ_{nm} and covariance matrix Σ_{nm}, which is assumed to be diagonal,

$$\mathcal{N}_{nm}(\mathrm{x_t}) = \frac{1}{(2\pi)^{D/2}|\Sigma_{nm}|^{1/2}} e^{-\frac{1}{2}(\mathrm{x}_t-\mu_{nm})^T \sum_{nm}^{-1}(\mathrm{x}_t-\mu_{nm})} \tag{5}$$

An enrollment speaker GMM is adapted from UBM parameters $\{\omega_{0m}, \mu_{0m}, \Sigma_{0m}\}$ using the MAP algorithm with sufficient training data $\mathrm{X}_n = \{\mathrm{x}_{n1}, \mathrm{x}_{n2}, \ldots, \mathrm{x}_{nTn}\}$. Here, the best MAP adaptation is achieved by adapting only the mean of the Gaussian components. The mean $\hat{\mu}_{nm}$ of the mth component of the Λ_n is updated via the following formula:

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

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

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

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

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

The speaker dependent model trained from a MAP-adapted UBM incorporates the speaker traits of the UBM. The speaker independent UBM covers only acoustic holes using sufficient training data, but the generalized coverage may damage the speaker dependent traits using the limited amount of enrollment speaker data with MAP adaptation.

C. Baseline GMM-cohort UBM system

Cohort-based speaker modeling as proposed by Prakash and Hansen (2007) is designed to mitigate the sparse phone coverage of limited enrollment data. The cohort speaker model assumes that acoustically similar speakers produce phones in a similar way. Here, acoustically similar speaker data are pooled to build the GMM, which consists of the in-set speaker cohort set. The in-set speaker cohort GMM is adapted via the MAP algorithm using the limited size enrollment data. The resulting GMM represents acoustically similar speakers, and this GMM will be more representative of the particular in-set speaker than a general UBM, which is constructed from a larger number of speakers to represent a complete anti-speaker model. The richer acoustic phoneme data from the selected cohort set are more effective in filling acoustic holes caused from limited amounts of enrollment data.

The cohort based in-set speaker model contains fewer speakers (with speaker n; $1 \leq n \leq N$) than the general UBM. It is necessary to choose the proper cohort speakers from the available development speaker pool. Selecting similar speakers who are acoustically close to the in-set speaker can be achieved in different ways; however, in this study, the baseline cohort system used by Prakash and Hansen (2007) is employed [GMM-Cohort UBM].

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

$$S_i = p(\mathbf{X}|\Lambda_i^{\mathrm{dev}}), \quad 1 \leq i \leq N_{\mathrm{dev}}. \quad (10)$$

(3) 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 in-set speaker n.
(4) Pool together the data of the selected cohorts and construct a cohort GMM as $\Lambda_n^{\mathrm{cohort}}$ for in-set speaker n.
(5) Using $\Lambda_n^{\mathrm{cohort}}$ as an initial model for the mean, covariance, and mixture weights, build the in-set speaker model Λ_n with the MAP algorithm.
(6) Repeat this procedure for each in-set speaker.

The GMM-cohort UBM illustrates the idea of filling acoustic holes for sparse enrollment data. The cohort speaker model assumes that acoustically close speaker data can be used to mitigate the limited enrollment data. Sections IV and V contribute to improve the system performance over the cohort baseline algorithm by choosing acoustically close speakers in a deliberate way. Instead of using *all the data from the cohort speaker,* we control the amount of data to build the in-set speaker model to match the phoneme distribution and density of the real speaker model based on the long conversational phoneme occurrence scenario from Fig. 1.

D. Baseline Eigenvoice system

The factor analysis based Eigenvoice algorithm is employed as the baseline system in this study (Kenny *et al.*, 2007, 2005). Speaker model, $\mathrm{M}(s)$, is defined by

$$\mathrm{M}(s) = \mathrm{M}_0 + \mathrm{V}\mathrm{y}(s). \quad (11)$$

Here, M_0 represents the supervector of the means of the UBM mixtures. The $\mathrm{y}(s)$ is a normal distributed speaker factor, and V is the Eigenvoice matrix, referred to as the total variability matrix (Kenny *et al.*, 2005).

The NIST 2004, 2005, 2006 SRE enrollment data are used to train a gender-dependent UBM with 1024 mixtures. The total variability matrix is trained on the Switchboard II Phases 2 and 3, Switchboard Cellular Part 1 and 2, and the NIST SRE 2004, 2005, and 2006 male enrollment data with five or more recording sessions per speaker. A total of 300 factors is selected for the Eigenvoice model. This represents the baseline Eigenvoice model.

III. GAUSSIAN MIXTURE MODEL MIXTURE TAGGING (GMT)

To compensate for limited enrollment data, two novel speaker adaptation algorithms are developed in this study, where both require a frame-level tagging method to be applied in these algorithms. By tagging acoustic information based on the speech feature sequence, it is possible to select the specific acoustic token data needed to fill holes. Phone recognition tagging using HMMs could be a first option to access the phone sequences of speech data. However, even the best phone recognizers obtain phone accuracies of the order of 65% ~ 75% (Lee and Hon, 1989; Lamel and Gauvain, 1993). Because there is up to 35% phone accuracy error, acoustic hole filling will be degraded by the direct phone classification step. To minimize this intrinsic phone recognition error, an alternate option is to train a GMM to classify acoustic information, and use GMT.

As a prior condition for GMT, the speaker independent Gaussian mixture model (S.I. GMM) should be built to classify acoustic information. Here, the entire available speaker set (total 438 speakers) is used to build the S.I. GMM, and it is noted that a large number of speaker based information will ultimately blend together for combined acoustic information in the resulting models. In Fig. 4, the procedure for building the S.I. GMM and classifying speaker information at the feature level is shown. The GMM contains a total of 438 speakers, which consist of 318 available cohort

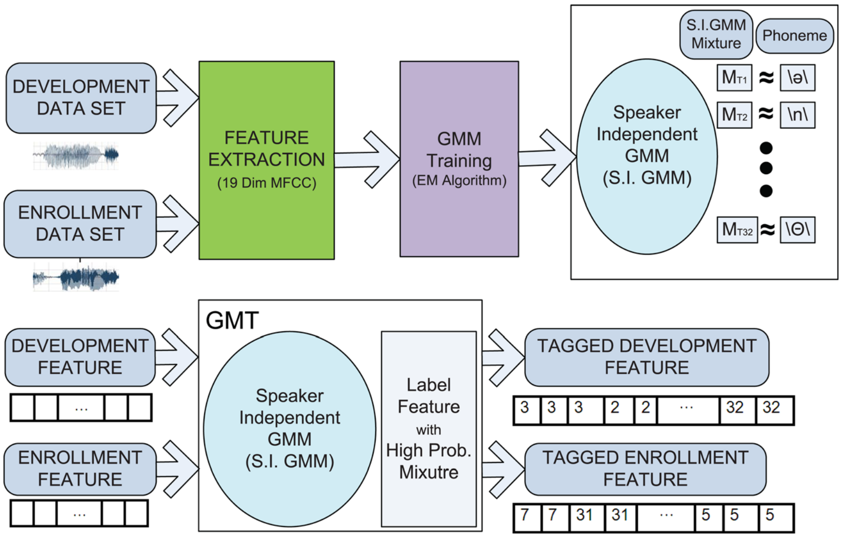

FIG. 4. (Color online) Block diagram of speaker independent (S.I.) GMM training and mixture tagging of speech features using the S.I. GMM.

speakers, 60 development speakers, and 60 in-set/out-of-set speakers. Because there is a low GMM mixture size, the resulting GMM can also be treated as a vector quantization process. The entire speaker data is used to train the GMM, so the GMM will perform better classification on the training data from which the model was just trained. It is suggested that the Gaussian components of the S.I. GMM will correspond to the acoustic space and therefore represent the acoustic tokens from a speaker independent acoustic perspective (Xiang and Berger, 2003; Ma *et al.*, 2006).

After obtaining the S.I. GMM, each feature frame of the speaker training data is labeled with the highest probability Gaussian mixture index for this GMT method, as shown in the lower part of Fig. 4. The speaker feature frame is classified into one Gaussian mixture index in the GMT scheme, and this tagging procedure is applied to both the enrollment and development data. The tagged information from the enrollment data is used for the first proposed speaker adaptation procedure, which uses different adaptation data depending on the ranking of tagged information (i.e., adaptation is directed toward mixtures that are represented by too little data from the original enrollment training speaker). Because this method needs ranking information for better adaption, tagging and the tagged number must be compared to make the necessary ranking. For the second speaker adaptation procedure, a GMT is developed for use in a balanced feature frame selection process. The second method uses a balanced table that contains the information on how much data are required for each Gaussian mixture tagged data to fill acoustic holes of the enrollment speaker data, in other words, how much enrollment data are required to achieve effective training performance seen for 10 min of data. Finding an optimum amount of data for the balanced acoustic space means the system will avoid borrowing unnecessary data for pdf mixtures with enough data and is important because too much outside cohort data will reduce the "dependent" value of the existing in-set sparse data set (e.g., the fact is, no cohort speaker is a perfect speaker match, so only the required minimum amount should be used to achieve the best overall performance).

IV. SPEAKER SIMILARITY MEASUREMENT

To fill acoustic holes for sparse training data, the enrollment speaker borrows data from acoustically close speakers. When acoustically similar data are selected from cohort speaker data, caution should be exercised to ensure a minimum number of cohorts for filling acoustic holes. Three methods are proposed to select similar characteristic speakers. The first uses the probability of the feature frame against the speaker model (Prakash and Hansen, 2007). The second uses the two speakers' statistical models along with the Kullback–Leibler (KL) divergence (Kullback and Leibler, 1951; Wu *et al.*, 2007). Finally, the third uses a prime distance manner to compare the two speakers' feature information.

Probability-score-method (PS-M): The first measurement scheme considers the probability scores of a potential cohort speaker model using enrollment data X_n to select the cohorts (Prakash and Hansen, 2007). This approach measures how close the enrollment speaker data are to the present in-set speaker model based on Eq. (10). Comparing enrollment speaker data to the speaker model is an indirect way to measure similarity because this method does not compare each speaker's feature data directly. The similarity measurement can be improved by a symmetrical measurement of both speaker traits,

$$S_{ni} = p(\mathrm{X}_n^{\text{enrollment}}|\Lambda_i^{\text{dev}}) + p(\mathrm{X}_i^{\text{dev}}|\Lambda_n^{\text{enrollment}}), \qquad (12)$$

where $1 \le i \le N_{\text{dev}}$ and $1 \le n \le N_{\text{enrollment}}$.

In Eq. (12), the second probability term collects the information on how close the potential cohort speaker data are to the in-set enrollment speaker model. The symmetrical scores assess the similarity between enrollment speaker and potential cohort speakers from the development data.

KL-score method (KL-M): The second approach for measuring speaker similarity uses the KL divergence score between two speaker models (Wu *et al.*, 2007), termed KL-M. This is accomplished by comparing two speakers at the speaker model level. It is noted that a distribution measurement using the KL is an unfair way to compare different amounts of data for each speaker model because the development speaker group has an unlimited amount of speech, but the enrollment speaker has only a limited amount (5 s) of data. The speaker model using two different sizes of data is not a fair model comparison unless both data sets are equally distributed. Here the GMT from Sec. III is used to ensure the same amount of enrollment data are selected for each of the cohort speakers. The procedure is as follows:

Step 1: Build mixture tag (M_{Ti}) histogram of short duration (5 s) available training data for each enrollment speaker.

Step 2: Select the potential cohort data to match the enrollment speaker histogram from Step 1.

a) Use mixture tagger to tag all data for a potential cohort speaker (318 potential development speakers).
b) Select mixture tagged frames from each potential cohort speaker data to match the mixture tag histogram from Step 1, (This ensures a consistent acoustic representation for the input speaker and each potential cohort speaker).
c) Move to Step 3 for training, Step 4 for distance measurement.

Step 3: Build the GMM for the enrollment speaker and potential cohort speakers.

a) Build the GMM with EM algorithm for the given enrollment speaker using 5 s of training data.
b) Using data from each potential cohort speaker, which has been matched to the mixture tag histogram (i.e., 5 s of data), build a GMM to test for cohort distance.

Step 4: Measure the distance between enrollment speaker and potential cohort speakers.

a) Find speaker distance between enrollment speaker and potential cohort speakers.
b) Repeat for all development cohort speakers (318 in our evaluation).
c) Select the top number of cohort speakers so that the closest speakers are used first, and only the minimum number of mixture tag entries that require hole filing data are used (e.g., reduces unwanted blending of cohort speakers when input data are sufficient).

Cepstral feature method (CF-M): The third approach is the most primitive strategy to find similar speakers by measuring enrollment and development data at the feature frame level, X_n and X_i, $1 \le n \le N_{\text{enrollment}}$ and $1 \le i \le N_{\text{dev}}$, which will be called the CF-M. The symmetrical measurement method is used to measure the cepstral difference between the enrollment and development features in a manner similar to Eq. (12), as follows:

$$S_{ni} = \sum_{k=1}^{T_n} \underset{1 \le p \le T_m}{\operatorname{argmin}}\, d(\mathrm{x}_{nk}, \mathrm{x}_{ip}) + \sum_{z=1}^{T_m} \underset{1 \le q \le T_n}{\operatorname{argmin}}\, d(\mathrm{x}_{iz}, \mathrm{x}_{nq})$$
$$1 \le k \le T_n \; k^{th} \text{ token for speaker } n$$
$$1 \le p \le T_m \; p^{th} \text{ token for speaker } i, \qquad (13)$$

where $d(\mathrm{x}_{nk}, \mathrm{x}_{ip}) = \sqrt{(\mathrm{x}_{nk}, \mathrm{x}_{ip})^2}$.

This procedure is a direct pairwise approach for comparing both speakers and requires an exhaustive computing time. The feature level comparison results in finding the most acoustically close speaker set among the three methods; however, it requires more computation time than any other method.

V. SPEAKER MODELING ALGORITHM

As an adaptation method for speaker modeling, MAP is selected based on the resulting experiments in this section. The experiment was performed to measure speaker recognition error rate using the two well-known methods, maximum likelihood (ML) and MAP, with progressively increasing enrollment data-set sizes. The question for selecting a speaker model construction algorithm given the data is as follows; when do we employ GMM training using ML (ML-GMM) versus MAP model adaptation from a UBM (MAP-GMM)? It is known that MAP estimation compensates for inaccurate estimates of ML in sparse enrollment data, and MAP estimates outperform ML in the sparse training data (Gauvain and Lee, 1994). For another case, when sufficient enrollment data for training is available, a trained model (ML-GMM) should perform better than one which uses model adaptation (MAP-GMM); however, with insufficient input training data, a trained model will have acoustic holes. Figure 5 shows that ML outperforms MAP with a crossover points at 15 s of data for the task of in-set/out-of-set speaker recognition. Loosely speaking, ML methods are able to estimate an accurate speaker dependent acoustic space better than a MAP estimate because a maximized likelihood of ML with sufficient data is more likely to cover the acoustic space for the input test data than a translation mean of the pdfs from a UBM. It should be noted that the starting EER value for MAP adaptation in Fig. 5 is conditioned on the quality of the initial UBM speaker model (e.g., better starting UBM will produce better results than a UBM trained with significantly distant speakers). If this model is initially trained with a very divergent/different set of speakers than the present target speaker, then greater amounts of adaptation data would be needed for MAP to achieve acceptable performance. This variability in performance is one reason effective cohort speaker selection is needed to build a good starting

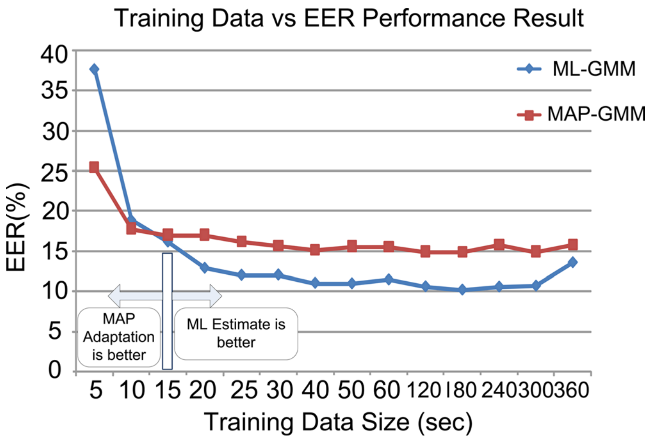

FIG. 5. (Color online) ML vs. MAP GMM training results EER performance is shown for each vs. available training data (6 sec test utterance).

base/UBM model for consistent MAP adaptation. In this study, however, the enrollment data are constrained to 5 s, and therefore MAP adaptation utilizing the short enrollment data is better than the ML method. Even if 10 s of enrollment data are used, Fig. 5 shows better performance using MAP versus ML. It corresponds to a previous finding that confirms MAP adaptation to be more suitable to sparse data than ML adaptation (Gauvain and Lee, 1994). The best performance, or in other words, the least error rate in Fig. 5 is near 10% equal error rate with sufficient data. This is an upper bound and an ideal EER goal, but it is noted that this 10% error rate is achieved with 180 s enrollment data set. Because this study uses only 5 s enrollment data, it is meaningful to have a better error rate than the least error rate with 5 s data, which is 25%.

To obtain better performance using MAP, selection of a speaker group for each enrolled speaker is important. A speaker model represents better enrolled speaker traits when MAP adaptation is performed with an acoustically similar cohort speakers' data than simply using random speaker adaptation data such as a UBM (Prakash and Hansen, 2007). This study extends this idea (Prakash and Hansen, 2007) by emphasizing and balancing the feature data using the GMT. First, the speaker model is constructed by emphasizing "sufficient data" and "acoustic hole" parts. The first speaker modeling confirms the results of filling acoustic holes by emphasizing/balancing the two parts of data. Second, we expand this idea by filling acoustic holes directly and the use a "balanced acoustic token histogram" via the GMT presented in Sec. III.

A. Speaker modeling: Top-down bottom-up procedure

The basic procedure begins with counting the number of mixture tagged tokens within the limited enrollment data. Because the amount of training data (e.g., 5 s) may not contain all 32 acoustic symbols, the procedure concentrates on identifying the most frequently occurring classes from top and in the least frequently occurring group from bottom. The most frequently occurring classes' cohort data are used to build the GMM called the "top" using EM algorithm, and the least frequently occurring classes' cohort data also built and called as the "bottom." Next, MAP adaptation fortifies those pdf mixtures for which the enrolled speaker's data are in the abundant top. The bottom class of feature data is also employed to move the means of the bottom GMM. This will be called as the "top-down bottom-up (TDBU)," where a comparison is made between the unsupervised MAP approach versus the supervised MAP approach. The TDBU speaker selection modeling method consists of the following steps, shown as a block diagram in Fig. 6: As a pre-step before applying TDBU procedure, the enrollment and development speech data feature frames are labeled with a 32 mixture class GMM using the GMT procedure from Sec. III, and the most acoustically similar set of speakers for each enrollment speaker n, $1 \leq n \leq N$ are selected (assume that one of the speaker similarity measures from Sec. IV has previously been applied).

Step 1: Count the most frequently occurring acoustic tokens (top) using the GMT, and the least occurring classes (bottom) for each enrollment speaker (i.e., organize the GMM pdfs in a top-down and bottom-up manner).

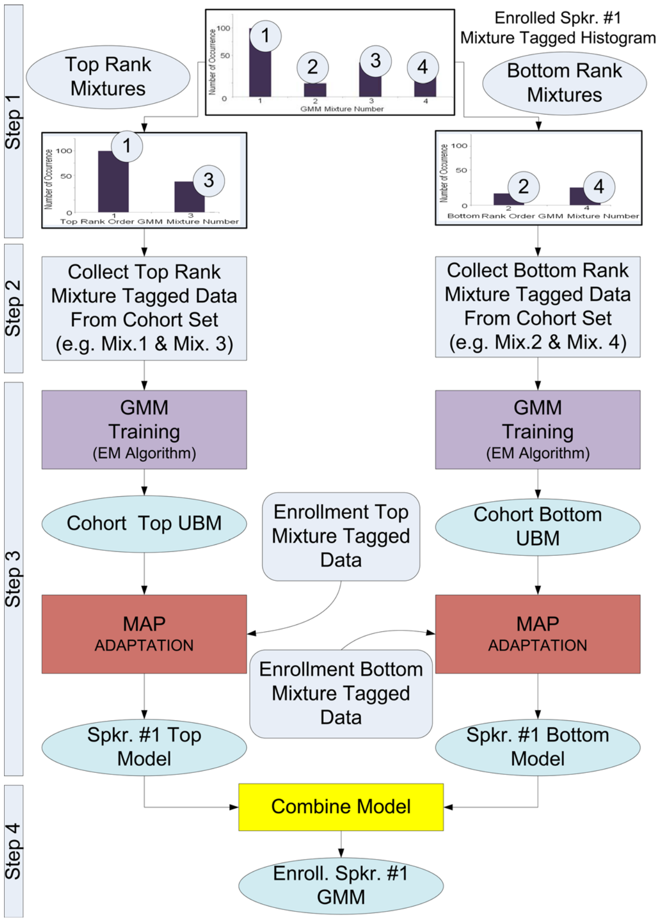
Step 1
Step 2
Step 3
Step 4
Top Rank Mixtures
Enrolled Spkr. #1 Mixture Tagged Histogram
Bottom Rank Mixtures
Number of Occurrence
100
50
0
1
2
3
4
GMM Mixture Number
Top Rank Order GMM Mixture Number
Bottom Rank Order GMM Mixture Number
Collect Top Rank Mixture Tagged Data From Cohort Set (e.g. Mix.1 & Mix. 3)
Collect Bottom Rank Mixture Tagged Data From Cohort Set (e.g. Mix.2 & Mix. 4)
GMM Training (EM Algorithm)
GMM Training (EM Algorithm)
Cohort Top UBM
Enrollment Top Mixture Tagged Data
Cohort Bottom UBM
MAP ADAPTATION
MAP ADAPTATION
Enrollment Bottom Mixture Tagged Data
Spkr. #1 Top Model
Spkr. #1 Bottom Model
Combine Model
Enroll. Spkr. #1 GMM

FIG. 6. (Color online) Block diagram of the TDBU enrolled speaker modeling process. Probe speaker modeling (TDBU) represents step by step procedure to build speaker dependent model.

Step 2: Collect the top and bottom data from the selected cohorts and construct a cohort GMM using EM algorithm as $\Lambda_n^{\text{top}-\text{cohort}}$ and $\Lambda_n^{\text{bottom}-\text{cohort}}$ for enrollment speaker n.

Step 3: Construct the enrollment speaker model Λ_n^{top} and $\Lambda_n^{\text{bottom}}$ with the corresponding top and bottom speaker data using MAP adaptation from the initial model, $\Lambda_n^{\text{top}-\text{cohort}}$ and $\Lambda_n^{\text{bottom}-\text{cohort}}$.

Step 4: Combine Λ_n^{top} and $\Lambda_n^{\text{bottom}}$ to build the final enrollment speaker model.

The top cohort class model $\Lambda_n^{\text{top}-\text{cohort}}$ is adapted using a relatively large amount of "top" enrollment speaker data via MAP, noting that the adapted top model Λ_n^{top} will have better discriminative speaker traits. The bottom class, $\Lambda_n^{\text{bottom}-\text{cohort}}$, can move the corresponding Gaussian components with the bottom class based enrollment data using MAP. The final enrollment speaker model combines Gaussian components from both top/bottom MAP models, and the weights of both GMMs are normalized by constants such as 0.7 for the top model weights and 0.3 for the bottom model weights for the experiment. This speaker modeling process probes the advantage of filling acoustic holes by finer separation of top and bottom parts of acoustic space. The next method focuses on which kind of data from the cohort models should be used to fill in the acoustic holes.

B. Mixture representation of natural conversation (MRNC) based cohort speaker modeling

Conversational English does have a different occurrence frequency of each phoneme (Mines *et al.*, 1978). The speaker model can be a far better representation when sufficient speech data covers all phoneme traits of the speaker for an input test utterance. A limited amount of enrollment data can be supplemented by determining how much data are needed to fill phoneme acoustic holes using the average phoneme occurrence from natural conversational speech. The average phoneme occurrence is counted as a unit in the indexed mixture of each speech feature frame using the GMT. A balanced acoustic token histogram is formed from this information, and is called the MRNC table. Each mixture indexed feature datum is collected from the acoustically similar cohort speaker set, so the number of used cohort speakers will depend on the required mixture amount in the MRNC table. The MRNC cohort speaker model is formed and adapted using the available enrollment speaker data. The MRNC speaker modeling procedure is as follows (and shown in Fig. 7):

Step 1: Count the average number of occurrences of each mixture tagged data for all development speaker's data i, $1 \le i \le N_{\text{dev}}$, using the GMT. With this, the MRNC table is constructed.

Step 2: Collect an equivalent number of the same mixture token feature frames using the MRNC table. The number of top N_{cohort} speakers for each enrollment speaker varies depending on how much filling data are required from the MRNC table and how much cohort filling data are available for each mixture class feature.

Step 3: Collect the data from the selected cohorts and construct an MRNC cohort GMM as $\Lambda_n^{\text{cohort}}$ for enrollment speaker n.

Step 4: Obtain the final enrollment speaker model Λ_n using MAP adaptation with the available (5 s) of enrollment speaker data from an initial model, $\Lambda_n^{\text{cohort}}$.

Filling acoustic holes using a cohort speaker model requires the development speakers' data to cover the vacant phone space. The MRNC table justifies how much data are needed for each mixture to fill the hole by representing the amount for each mixture in the conversational speech, and the final MRNC cohort model achieves acoustic hole filling by effective selection of cohort speakers from which the hole filling data are drawn.

VI. EXPERIMENTS

A. Experimental setup

The proposed algorithms are evaluated using the FISHER (Cieri *et al.*, 2004) and CU-MOVE (?). Speaker recognition employing telephone speech is a primary research task in the speech community because there is often a need to verify the identity of a caller without visual information. The parameters of EM/MAP methods are fixed to compare the various algorithm performances.

1. *FISHER corpus*

An in-set/out-of-set speaker group is defined from the FISHER corpus. In a similar manner to Angkititrakul and Hansen (2007) and Prakash and Hansen (2007), an in-set/out-of-set group of speakers is randomly selected. Here, 60 male speakers are drawn as members of in-set/out-of-set groups. For experimentation, three in-set/out-of-set groups are formed that include 15 in-set/45 out-of-set, 30 in-set/30

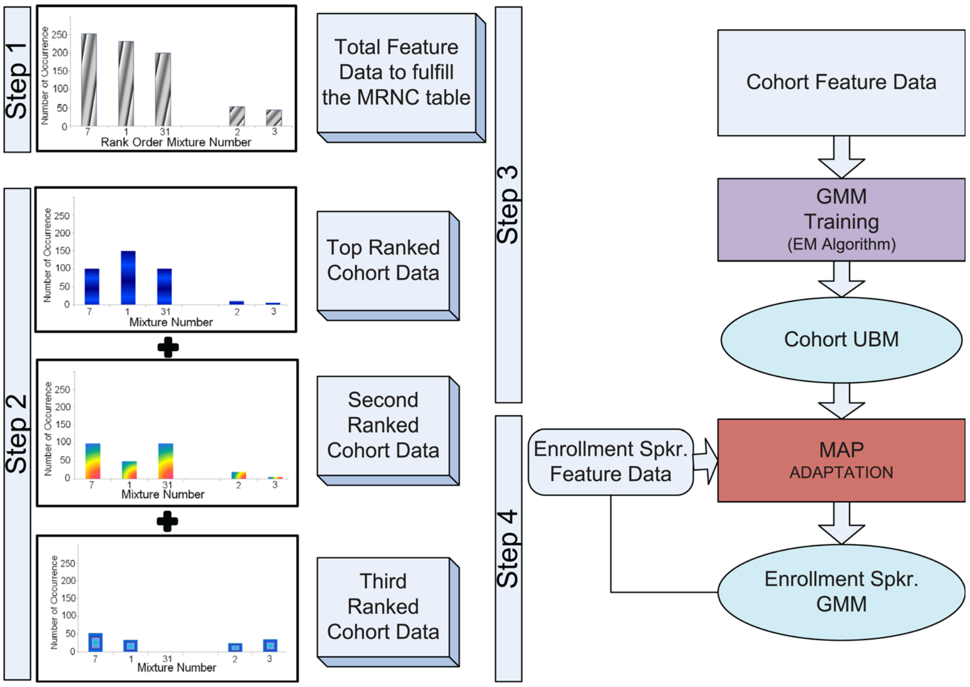

FIG. 7. (Color online) Block diagram shows cohort data collected from the GMT tagged features by imitating the phoneme-like distribution in the mixture representation of natural conversation (MRNC) table. The final speaker model is adapted from the MRNC cohort model.

out-of-set, and 45 in-set/15 out-of-set speakers. All 60 speakers are devoted to the in-set or out-of-set groups with 50 randomly chosen combinations for three different conditions. The in-set speaker group does not overlap with any speaker from the out-of-set group, so all 60 speakers are present in one of either set. Training data for the enrollment speaker are about 5 s of speech, and test utterance duration per speaker is 2, 4, and 6 s audio utterances. The development speaker corpus consists of 378 speakers, who have about 30 s of speech data in each.

2. CU-MOVE corpus

An in-set/out-of-set speaker group using the CU-MOVE corpus is defined in a similar manner as that used for the FISHER corpus. The CU-MOVE speech corpus consists of speech and noise data collected from across the USA within vehicles under a variety of driving conditions to facilitate design of in-vehicle interactive systems for route planning and navigation (Hansen *et al.*, 2000). In this study, the spontaneous speech part of the corpus is used to evaluate the proposed algorithms. Test data durations consist of 2, 4, and 6 s of speech. Ninety-six male speakers are used as development data.

3. Feature preparation

The GMM-MAP based system and Eigenvoice solution employ the same Mel-frequency cepstral coefficients (MFCC), but both systems use a different dimensional feature set. For the GMM-MAP system, the extracted sequence of feature frames use a 20 ms analysis window length with a 10 ms skip rate applied for consecutive frames. Each frame consists of 19-dim static MFCC with silence and low energy speech frames removed using a general frame energy detection technique.

For Eigenvoice parameterization, a 60-dimensional feature set (19 MFCC with log energy plus delta (Δ) and delta-delta ($\Delta\Delta$) is used, where features are extracted using a 25 ms analysis window with 10 ms skip rate, filtered by feature warping (Pelecanos and Sridharan, 2001) using a 3 s sliding window. Voice activity detection (VAD) is applied to remove low energy speech frames based on an energy threshold. Here, the test audio segment is shorter than the feature warping sliding window size, and audio segment of 2/4/6 s is combined for parameterization, and then feature is divided into the same portion of 2/4/6 s. For example, 2/12 portion of extracted feature is used for 2 s of test data.

B. Speaker similarity measurement

The cohort-based speaker model procedure begins with borrowing speech data from the cohort set to cover acoustic holes. The set of cohort speakers plays a crucial role in properly filling acoustic holes. The performance of probability measurement score used in the previous study (Prakash and Hansen, 2007) is improved by adding the probability score of the development data given the in-set speaker model as shown in Eq. (12). The cohort baseline experiment here uses the combined probability scores. The KL-M is performed next. The KL requires both an in-set speaker model and a potential cohort development speaker model, where both models are constrained to have the same amount of data. The in-set speaker has about 5 s of data, so the development speaker's data is collected to ensure an equivalent amount with the in-set speaker. We build the 16 Gaussian components of a GMM for both models using EM algorithm. The third cepstral feature comparison is performed in the last experiment (CF-M). Again, 19-dim MFCC features for in-set and potential cohort speaker's data are compared. In a manner similar to the probability score measurement, the cepstral measurement calculates the in-set to potential cohort speaker data space as shown in Eq. (13). This primitive measure which takes direct feature-to-feature distance requires an extensive amount of computation cycles. The EER results for in-set/out-of-set speaker recognition are shown in Table I, and indicate that cepstral speaker similarity measure (CF-M) performs the best among the three measures. The word GMM is omitted in Table I for cohort UBM baseline. The CF-M improves an average of 10.8% relative EER over the GMM-cohort UBM baseline (Prakash and Hansen, 2007). KL measurement showed better results than GMM-cohort UBM baseline and the equal amount of data for the KL measurement is a key reason why there is an average 4.3% relative EER improvement over the previous GMM-cohort UBM baseline. The total number of cohorts used in each measure is 5 for the GMM-cohort UBM and KL-M Cohort, and 6 for the CF-M cohort.

TABLE I. Speaker similarity performance [in terms of EER(%)] using 6 s of test data for *in-set*/*out-of-set* speaker.

	EER		
	15 in/45 out	30 in/30 out	45 in/15 out
GMM-UBM baseline	23.98	25.20	26.58
Cohort UBM baseline	19.84	24.43	24.75
Cohort UBM (KL-M)	19.16	22.73	24.11
Cohort UBM (CF-M)	17.75	21.27	22.56

C. Speaker modeling experiment: TDBU

The TDBU speaker modeling explores the prospects of using a short amount of enrollment to improve overall system performance. Table II shows GMM-cohort UBM baseline results that use KL-M for selecting the cohorts. The best performance of TDBU is obtained using KL-M. The general UBM consists of 32 Gaussian components of a GMM built

TABLE II. Comparison performance of GMM-UBM baseline, GMM-cohort UBM baseline, and TDBU speaker modeling using KL-M with 6 s of test utterance [EER(%)].

	EER		
	15 in/45 out	30 in/30 out	45 in/15 out
GMM-UBM baseline	23.98	25.20	26.58
GMM-Cohort UBM	19.16	22.73	24.11
TDBU	18.68	22.60	23.48

with 30 s of data from each of 60 development speakers (30 min of UBM data). In a manner presented in Sec. V A, the top cohort model consists of 16 Gaussian components and the bottom also has the same number of Gaussian. The number of Gaussian components for top/bottom model is selected based on the average amount of training data (e.g., 85% of the 5 s of training data presents top 16 model, and 15% represents bottom model). Both models are combined with a fixed 7:3 weight ratio (all mixture weights for Λ_n^{top} with 0.7, all mixture weights for $\Lambda_n^{\text{bottom}}$ by 0.3), and the weight ratio is selected based on heuristic results. Test utterances are all 6 s in duration. The KL speaker similarity is performed for both the GMM-cohort UBM baseline and TDBU cohort modeling.

The results shown in Table II confirms that the TDBU method improves EER performance than other methods. TDBU speaker modeling emphasizes the importance of adaptation for short amounts of enrollment data. Here, the top cohort model has better discrimination, and adaptation with about 85% of the enrollment data, and the bottom cohorts migrate the Gaussians to the in-set speaker traits with about 15% of the enrollment data. Supervised MAP for both the top and bottom part of GMM shows a 1.90% relative improvement over the GMM-cohort UBM baseline using 5 cohort speakers for both methods. The relative improvements in EER varies from 2.6% ∼ 5.3% for 15/45, 30/30, and 45/15 in-set/out-of-set speaker configurations over the GMM-UBM baseline.

D. MRNC cohort speaker modeling

In this section, the new MRNC cohort-based speaker modeling is compared with the GMM-UBM baseline, GMM-cohort UBM, and Eigenvoice baseline. The GMM-cohort UBM baseline's cohort set is selected by speaker similarity assessment using the CF-M. MRNC table is calculated the same way both in FISHER and CU-MOVE data. Both in-set/out-of-set speaker models have 32 Gaussian components, and Eigenvoice sets the factor dimension as 300.

1. Evaluation FISHER corpus

The in-set/out-of-set consists of 60 speakers, and the UBM is trained with 60 development speakers. Each in-set speaker has access to 318 cohort development speakers who are ranked ordered with speaker similarity measures and who do not overlap with the 60 UBM development speakers. All development speakers have about 30 s of speech. Each in-set speaker has about 5 s of training data, and evaluation performed with 2, 4, or 6 s test utterances. Figure 8 summarizes results, showing the best results obtained with 60 c of data distributed based on the MRNC table. The GMM-cohort UBM baseline employs five cohort speakers, which collectively have about 150 s data. As the in-set size increases, overall system performance decreases, with a corresponding increase in EER. Longer test duration data produces improvement over the 2 s short test utterance. Absolute EER improvement of MRNC cohort modeling range from 1.79% ∼ 7.38%, and an average 16% relative improvement in EER is achieved for 15/45, 30/30, and 45/15 over the GMM-UBM baseline. The proposed MRNC produces an average 7.42% relative improvement over GMM-cohort UBM baseline and 19% relative improvement over Eigenvoice.

The performance of MRNC method is comparable to GMM-UBM baseline in respect to enrollment data size compensation scheme. The system performance using 60 s speaker data via GMM-UBM adaptation is comparable to 5 s speaker data using MRNC model by comparing Figs. 5 and 8. While the size of in-set/out-of-set is different by 15/15 in Fig. 5. and 15/45 and 30/30 in Fig. 8, the test size is same in both experiments. GMM-UBM trained 60 s or 360 s model results 15% EER, and MRNC method results in a 16.1% and 18.7% EER on 6 s test data for 15/45 and 30/30 in-set/out-of-set. MRNC method using only 5 s drops in performance by 1.1% EER versus the GMM-UBM using a 60 s for the same in-set speaker size. The MRNC acoustic hole filling algorithm produces similar performance to that seen for the GMM-UBM which is trained with 60 s of data, which is 12 times as much data.

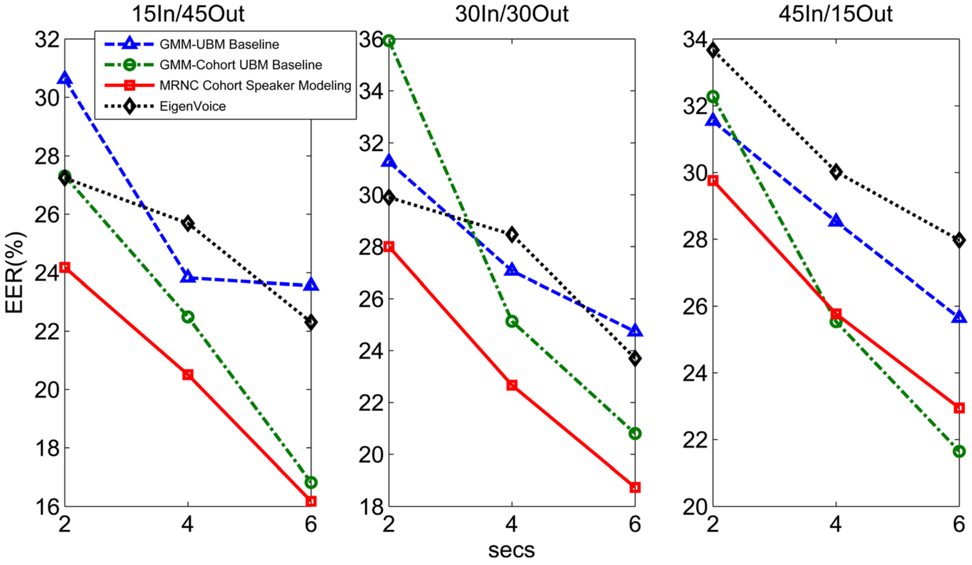

FIG. 8. (Color online) Performance[in terms of EER(%)] of in-set/out-of-set speaker recognition using FISHER Corpus with 2, 4, 6 s test utterances for GMM-UBM baseline, GMM-cohort UBM Baseline, Eigenvoice, and proposed MRNC cohort speaker modeling.

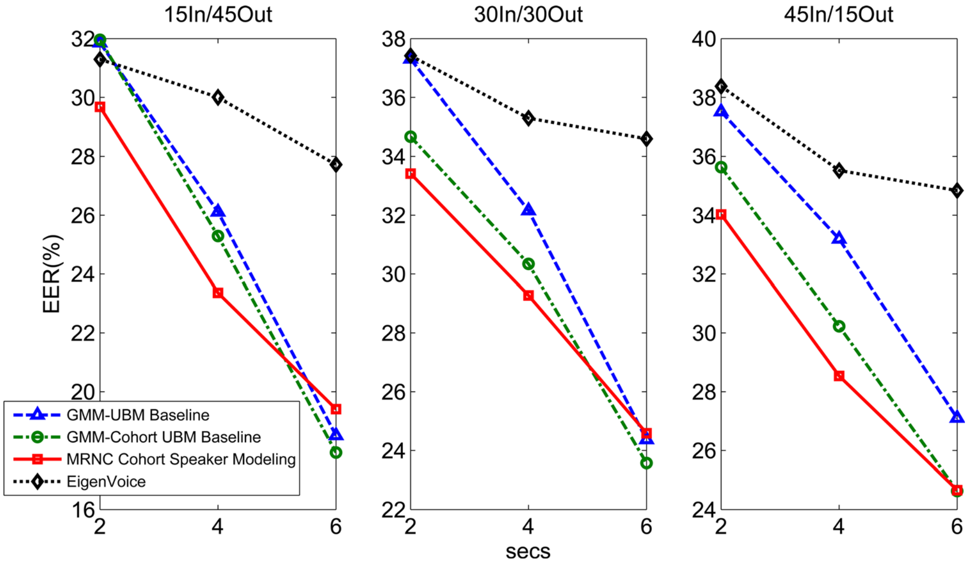

FIG. 9. (Color online) Performance[in terms of EER(%)] of in-set/out-of-set speaker recognition using CU-MOVE corpus with 2, 4, 6 s test utterances for GMM-UBM baseline, GMM-cohort UBM baseline, Eigenvoice, and proposed MRNC cohort speaker modeling.

2. Evaluation CU-MOVE corpus

For equal comparison between FISHER and CU-MOVE corpora, the GMM construction and adaptation parameters are set equal to both corpora. The in-set/out-of-set consists of 60 speakers, and the UBM is trained with 60 development speakers. Each in-set speaker has about 5 s of training data, and evaluation performed with 2, 4, or 6 s test utterances. All development speakers have about 60 s of speech. Figure 9 summarizes results, showing the best results obtained with 60 s of data distributed based on the MRNC table. The GMM-cohort UBM baseline employs eight cohort speakers, about 480 s data. Similar as the results using Fisher corpus, the in-set size increases, the performance decreases and longer test duration produces higher performance. The MRNC cohort model performs superior over GMM-UBM/GMM-cohort UBM in 2 and 4 s test data by average absolute EER 1.64% and 3.31% but degrades its discriminating ability in the 6 s test case. The proposed MRNC method is effective for extreme short test utterance. Eigenvoice for the noisy CU-MOVE corpus does not perform well because the mismatch training condition for Eigen-voice degrades the CU-MOVE evaluation performance. The CU-MOVE has been also down sampled to 8 kHz for Eigenvoice evaluation. These two effects degrade the performance of Eigenvoice, and it is prohibitive to use Eigenvoice for this corpus unless an effective car noise background corpus has been developed.

E. Human listener evaluation

Human speaker recognition evaluation is performed to study the benefit of in-set/out-of-set in human listening, and these results are compared with machine performance. It is noted that forensic based speaker recognition and human assisted speaker recognition (HASR) are emerging research areas. The latest NIST-SRE 2010 evaluation considered HASR with ∼5 min reference and test file (Schwartz *et al.*, 2010; Greenberg *et al.*, 2010), and a number of studies are beginning to explore the benefits of statistical/automatic speaker identification coupled with human speaker identification. Here the focus is simply to benchmark a human in-set/out-of-set speaker recognition test with an automatic scheme when limited data is available. Human listener evaluation here is based on the FISHER corpus, where the data size is 5 s for both the enrollment and test speakers. Listeners are asked to recognize 10/10 in-set/out-of-set speaker groups, where these speakers are selected from FISHER evaluation corpus. Each speaker has two speech tokens, where a total of 40 utterances streams are used for in-set/out-of-set speaker recognition. After listening to the 10 speech utterances for in-set speaker group, the listener must answer the question; "Does the speaker in this audio belong to the in-set speaker group?". If they choose "yes" for this question, then they are asked to select which target speaker this file belongs to among the in-set speaker group. Therefore, this task is an open-set speaker identification test. A total of 11 listeners participated in the evaluation of the in-set/out-of-set and open-set speaker recognition tasks.

The results here shown that human do not perform as well as machines in speaker recognition with sparse data. The machine has better performance in both in-set/out-of-set recognition with 95% accuracy compared to only 72.2% for human in-set/out-of-set performance. For the task of open-set speaker identification, the machine performance is 85%, while human performance is reduced to 68.8%. The GMM-UBM baseline algorithm is used for the machine based speaker recognition, and the machine recognition accuracy is measured using a threshold of 0 for a binary decision. The smaller in-set/out-of-set size and binary decision results higher accuracy for the machine. These results do not completely follow the findings in a previous study (Schmidt-Nielsen and Crystal, 2000), where human performance is better at speaker recognition than machine using speech degraded by background noise or various handsets. One reason why human performance is worse can be the limit of human "acoustic" memory. In this study, the number of in-set speakers is set to 10, so listeners would need to remember all 10 speakers to perform the test. Therefore, the performance represents the case where listeners are not familiar with the speakers (i.e., this performance would presumably change if speaker familiarity is greater for the listeners).

VII. DISCUSSION

This study has focused on the task of filling acoustic holes in the speaker GMM space due to very short enrollment data. A speaker similarity measurement is employed to select cohort speakers for proper filling based on new speaker modeling methods. The speaker model measure using KL-M improves the average relative EER by 4.3%, and the CF-M provides an average relative improvement of 6.7% over a previously developed GMM-cohort UBM baseline (Prakash and Hansen, 2007). The final proposed algorithm uses CF-M along with incorporating cohort speakers' with balanced acoustic data in a manner proportional to the natural phoneme occurrence without direct phoneme recognition knowledge (MRNC table). It is noted that a smaller in-set group performs better than if the in-set group is larger because a smaller in-set group has a larger pool of speaker to reject; greater confusion with open out-of-set speakers is possible when the in-set group size expands. The proposed system performs well on longer duration test material because the longer test utterances provides a larger sample of the acoustic space versus the speaker model. An average 16% relative improvement, and 1.79% ~ 7.38% absolute improvement is observed using the final proposed MRNC cohort algorithm with the FISHER corpus. The evaluation based on the in-vehicle environment speech corpus, CU-MOVE, also shows improvement over the baseline systems with short test data. The FISHER corpus was employed here versus NIST SRE-08 to have a consistent English language and common microphone to avoid channel variation. Again, the focus was to address the lack of phoneme space, and therefore the decision was made to set aside channel effects of the corpus domain. Even with this constraint on handset variability, channel compensation represents an additional goal for future work. For human listener evaluation, employing a smaller number of in-set speakers can reduce acoustic memory load, where it is possible to explore human listening ability to fill the acoustic holes by comparing human and machine performance.

VIII. CONCLUSION

This investigation has focused on the specific text independent speaker recognition problem of in-set/out-of-set with extremely short amounts of enrollment and test data. A method was proposed to classify speech feature frames into a 32 mixture index class using a GMT procedure. This phone-like classification allows analysis of the indexed mixture occurrence of longer duration conversation speech and to identify the acoustic tokens necessary to fill acoustic holes for the enrollment speaker space. Based on a previous developed GMM-cohort UBM baseline system (Prakash and Hansen, 2007), we proposed a method for selecting similar cohort speakers with each enrollment speaker. The measures included a probability score method (PS-M), an assessment between speaker models (KL-M), and finally a measure between MFCC features (CF-M). The cohort baseline shows significant improvement in performance when more acoustically similar related speaker groups are selected. The MRNC table is used to incorporate frame tagged data filling with cohort-based speaker models. The proposed MRNC cohort-based speaker models imitate the phone-like distribution of natural conversational speech, so the resulting model mimics the natural speaker conversation behavior for the input enrollment speaker. Finally, future work could employ channel compensation schemes so that a larger pool of cohorts from other corpora could be made available. In addition, it could be possible to build similar speaker groups for normalizing the test speakers or develop an alternative decision threshold for acoustically similar speakers. Human listener evaluation was performed to compare the speaker recognition performance with machine. It is hard to judge the speaker recognition performance between human and machine due to human ability for a given task. However, in this study, humans do not perform as well as machines for speaker recognition with extreme sparse data using unfamiliar in-set/out-of-set speakers. For future work, the MRNC cohort speaker model can be applied to Eigenvoice speaker adaptation to compensate the sparse enrollment data. Speaker ranking methods can be also studied for building cohort speaker groups.

ACKNOWLEDGMENTS

This project was funded by AFRL through a subcontract to RADC Inc. under FA8750-09-C-0067, and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by John H. L. Hansen.

Angkititrakul, P., and Hansen, J. H. L. (**2007**). “Discriminative in-set/out-of-set speaker recognition,” IEEE Trans. Audio, Speech, Lang. Process. **15**, 498–508.

Burges, C. (**1998**). “A tutorial on support vector machines for pattern recognition,” Data Min. Knowl. Discov. **2**, 121–167.

Cieri, C., Miller, D., and Walker, K. (**2004**). “The fisher corpus: A resource for the next generations of speech-to-text,” in Fourth International Conference on Language Resources and Evaluation May 2004, Lisbon, Portugal, Vol. 1, pp. 1–3.

Furui, S. (**1997**). *Recent Advances in Speaker Recognition* (Springer, Berlin, Germany), pp.235–252.

Gauvain, J., and Lee, C. (**1994**). “Maximum a posteriori estimation for multivariate Gaussian mixture observations of markov chains,” IEEE Trans. Speech Audio Process. **2**, 291–298.

Gish, H., and Schmidt, M. (**1994**). “Text-independent speaker identification,” IEEE Signal Process. Mag. **11**, 18–32.

Greenberg, G., Martin, A., Brandschain, L., Campbell, J., Cieri, C., Doddington, G., and Godfrey, J. (**2010**). “Human assisted speaker recognition in NIST SRE10,” in *Odyssey 2010*, Brno, Czech Republic, pp. 180–185.

Hansen, J. H. L., Huang, R., Zhou, B., Seadle, M., Deller, J., Gurijala, A., Kurimo, M., and Angkititrakul, P. (**2005**). “Speechfind: advances in spoken document retrieval for a national gallery of the spoken word,” IEEE Trans. Speech Audio Process. **13**, 712–730.

Hansen, J. H. L., Plucienkowski, J., Gallant, S., Pellom, B., and Ward, W. (**2000**). “CU-move: Robust speech processing for in-vehicle speech systems,” in *ICSLP 2000*, Beijing, China, pp. 524–527.

Huber, P. (**1965**). “A robust version of the probability ratio test,” Ann Math. Stat. **36**, 1753–1758.

Kenny, P., Boulianne, G., and Dumouchel, P. (**2005**). “Eigenvoice modeling with sparse training data,” IEEE Trans. Audio, Speech, Lang. Proc. **13**, 345–354.

Kenny, P., Boulianne, G., Ouellet, P., and Dumouchel, P. (**2007**). “Speaker and session variability in GMM-based speaker verification”, IEEE Trans. Audio, Speech, Lang. Process. **15**, 1448–1460.

Kressel, U., and Schurmann, J. (**1997**). *Pattern Classification Techniques Based on Function Approximation*, (World Scientific Publishing, MA), pp. 49–78.

Kuhn, R., Junqua, J., Nguyen, P., and Niedzielski, N. (**2000**). “Rapid speaker adaptation in Eigenvoice space,” IEEE Trans. Speech Audio Process. **8**, 695–707.

Kullback, S., and Leibler, R. (**1951**). “On information and sufficiency,” Ann. Math. Stat. **22**, 79–86.

Lamel, L., and Gauvain, J. (**1993**). “High performance speaker-independent phone recognition using CDHMM,” in *EUROSPEECH 1993*, Berlin, Germany, pp. 121–124.

Lee, K., and Hon, H. (**1989**). “Speaker-independent phone recognition using hidden Markov models,” IEEE Trans. Acoust., Speech, Signal Process. **37**, 1641–1648.

Li, Q., Juang, B., Lee, C., Zhou, Q., and Soong, F. (**2003**). *Speaker Authentication* (CRC Press, New York), pp. 229–259.

Ma, B., Zhu, D., Tong, R., and Li, H. (**2006**). “Speaker cluster based GMM tokenization for speaker recognition,” in *INTERSPEECH 2006*, Pittsburgh, PA, Vol. 1, pp. 505–508.

Mak, M., Hsiao, R., and Mak, B. (**2006**). “A comparison of various adaptation methods for speaker verification with limited enrollment data”, in *ICASSP 2006*, Toulouse, France, Vol. 1, pp. 929–932.

Matrouf, D., Bellot, O., Nocera, P., Linares, G., and Bonastre, J. (**2003**). “Structural linear model-space transformations for speaker adaptation,” in *EUROSPEECH 2003*, Geneva, Switzerland, pp. 1625–1628.

Mines, M., Hanson, B., and Shoup, J. (**1978**). “Frequency of occurrence of phonemes in conversational English” Lang. Speech **21**, 221–241.

Mitra, V., Garcia-Romero, D., and Espy-Wilson, C. (**2008**). “Language and genre detection in audio content analysis,” *INTERSPEECH-2008*, Brisbane, Australia, pp. 2506–2509.

Pelecanos, J., and Sridharan, S. (**2001**). “Feature warping for robust speaker verification,” Proc. Speaker Odyssey **13**, 1–5.

Prakash, V., and Hansen, J. H. L. (**2007**). “In-set/out-of-set speaker recognition under sparse enrollment,” IEEE Trans. Audio, Speech, Lang. Process. **15**, 2044–2052.

Reynolds, D., Quatieri, T., and Dunn, R. (**2000**). “Speaker verification using adapted gaussian mixture models,” Digit. Signal Process. **10**, 19–41.

Scheffer, N., and Bonastre, J. (**2006**). “A multiclass framework for speaker verification within an acoustic event sequence system,” in *INTERSPEECH 2006*, Pittsburgh, PA, pp. 501–504.

Schmidt-Nielsen, A., and Crystal, T. (**2000**). “Speaker verification by human listeners: Experiments comparing human and machine performance using the NIST 1998 Speaker Evaluation Data* 1,” Digit. Signal Process. **10**, 249–266.

Schwartz, R., Campbell, J., Shen, W., D., S., Campbell, W., Richardson, F., Dunn, R., and Granville, R. (**2010**). “USSS-MITLL 2010 Human Assisted Speaker Recognition Evaluation System,” in *NIST SRE Workshop 2010*, Brno, Czech Republic, pp. 1–7.

Sivakumaran, P., Fortuna, J., and Ariyaeeinia, A. (**2003**). “Score normalisation applied to open-set, text-independent speaker identification,” in *EUROSPEECH*, Geneva, Switzerland, pp. 2669–2672.

Soong, F., Rosenberg, A., Rabiner, L., and Juang, L. (**1985**). “A vector quantization approach to speaker recognition,” in *ICASSP 1985*, FL, pp. 387–390.

Torres-Carrasquillo, P., Singer, E., Kohler, M., Greene, R., Reynolds, D., and Deller, Jr., J. (**2002**). “Approaches to language identification using Gaussian mixture models and shifted delta cepstral features,” in *ICSLP 2002*, Tampa, FL, pp. 89–92.

Wu, W., Zheng, T., Xu, M., and Soong, F. (**2007**). “A cohort-based speaker model synthesis for mismatched channels in speaker verification,” IEEE Trans. Audio, Speech, Lang. Process. **15**, 1893–1903.

Xiang, B., and Berger, T. (**2003**). “Efficient text-independent speaker verification with structural gaussian mixture models and neural network,” IEEE Trans. Audio, Speech, Lang. Process. **11**, 447–456.