Improving Short Utterance Speaker Recognition by Modeling Speech Unit Classes

Lantian Li, Dong Wang, *Member, IEEE*, Chenhao Zhang, and Thomas Fang Zheng, *Senior Member, IEEE*

***Abstract*—Short utterance speaker recognition (SUSR) is highly challenging due to the limited enrollment and/or test data. We argue that the difficulty can be largely attributed to the mismatched prior distributions of the speech data used to train the universal background model (UBM) and those for enrollment and test. This paper presents a novel solution that distributes speech signals into a multitude of acoustic subregions that are defined by speech units, and models speakers within the subregions. To avoid data sparsity, a data-driven approach is proposed to cluster speech units into speech unit classes, based on which robust subregion models can be constructed. Further more, we propose a model synthesis approach based on maximum likelihood linear regression (MLLR) to deal with no-data speech unit classes. The experiments were conducted on a publicly available database SUD12. The results demonstrated that on a text-independent speaker recognition task where the test utterances are no longer than 2 seconds and mostly shorter than 0.5 seconds, the proposed subregion modeling offered a 21.51% relative reduction in equal error rate (EER), compared with the standard GMM-UBM baseline. In addition, with the model synthesis approach, the performance can be greatly improved in scenarios where no enrollment data are available for some speech unit classes.**

***Index Terms*—Short Utterance, Speaker Recognition, Subregion Model, Model Synthesis.**

I. Introduction

Speaker recognition aims to recognize claimed identities of speakers, including identification and verification. It has gained great popularity in a wide range of applications including access control, forensic evidence provision, and user authentication in telephone banking. After decades of research, current speaker recognition systems have achieved rather satisfactory performance, given that the enrollment and test utterances are sufficiently long and the signal-to-noise ratio (SNR) is large enough [1]–[5].

A traditional approach to speaker recognition is the GMM-UBM framework [6], [7]. This approach involves a Gaussian mixture model (GMM) based universal background model (UBM) to represent the probability distribution of acoustic features from all speakers, and each enrolled speaker is represented by a Gaussian mixture model (GMM) which is adapted from the UBM via maximum *a posteriori* (MAP) estimation [8].

Another main-stream approach is based on joint factor analysis (JFA) and its 'simplified' version, the so-called i-vector model. While JFA models speaker and channel variabilities in two separate subspaces [9], the i-vector approach models these variabilities in a single low-dimensional subspace [10].

To improve the i-vector model, a multitude of normalization techniques have been studied and employed, such as with-in class covariance normalization (WCCN) [11], [12], nuisance attribute projection (NAP) [2], [10] and probabilistic LDA (PLDA) [13]. These methods have been demonstrated to be highly successful [5].

Recently, deep learning has gained much success in multiple domains and caused extensive interests [14]. For speaker recognition, a very recent study applies DNN models trained for speech recognition to build UBMs, so that rich information in phones can be employed to construct more accurate background models [15], [16]. Additionally, DNNs has been utilized to extract speaker features [17], [18].

A. Challenge With Short Utterance

In spite of the great achievement, current speaker recognition systems perform well only if the enrollment and test data are sufficiently long. In many applications, however, users are reluctant to provide much speech data particularly at the test phase, for instance in telephone banking. In other situations, it is highly difficult to collect sufficient data, for example in forensic applications.

If the enrollment and test utterances contain the same phone sequence (so called 'text-dependent' task), short utterances would not be a big problem [19]; however for text-independent tasks, severe performance degradation is often observed if the enrollment/test utterances are not long enough, as has been reported in several previous studies.

For instance, Vogt et al. [20] reported that when the test speech was shortened from 20 seconds to 2 seconds, the performance degraded sharply in terms of equal error rate (EER) from 6.34% to 23.89% on a NIST SRE task. Mak et al. [21] showed that when the length of the test speech is less than 2 seconds, the EER raised to as high as 35.00%.

B. Research on Short Utterance Speaker Recognition

Research on short utterance speaker recognition (SUSR) is still limited. In [22], the authors show that performance on

Manuscript received May 05, 2015; revised December 31, 2015 and February 23, 2016; accepted March 12, 2016. Date of publication March 21, 2016; date of current version April 29, 2016. This work was supported in part by the National Natural Science Foundation of China under Grants 61371136 and 61271389, and in part by the National Basic Research Program (973 Program) of China under Grant 2013CB329302. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Rodrigo Capobianco Guido.

The authors are with the Center for Speech and Language Technologies, Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: fzheng@tsinghua.edu.cn).

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/TASLP.2016.2544660

short utterances can be improved through the JFA framework that models speaker and channel variabilities in two separate subspaces. This work is extended in [23] which reports that the i-vector model can distill speaker information in a more effective way so it is more suitable for SUSR. In addition, a score-based segment selection technique has been proposed in [24], which evaluates the reliability of each test speech segment based on a set of cohort models, and scores the test utterance with the reliable segments only. A relative EER reduction of 22% was reported by the authors on a recognition task where the test utterances are shorter than 15 seconds in length.

It should be noted that the results reported in these research studies are based on test utterances that are 5∼10 seconds long. This is still rather long in many scenarios. For very short test utterances, i.e., 1∼2 seconds in length, there are no satisfactory solutions yet, to the authors′ best knowledge. In addition, if the enrollment utterance is also short, the recognition will be more challenging, for which very little research has been conducted. This paper focuses on improving the recognition performance on very short test utterances where the valid speech is of 2 words in maximum, and dealing with the situation where both enrollment and test utterances are short.

C. Motivations

We argue that the difficulty associated with SUSR can be largely attributed to the mismatched distributions of the speech data used to train the UBM and to enroll/test a particular speaker. Following the standard framework of GMM-UBM, the characteristic of a particular speaker is modeled by a GMM. A commonly adopted GMM-UBM setup is to train a UBM on a pool of speech data involving a large number of speakers via the EM algorithm [25], and then a speaker's model is derived from the UBM given the enrollment speech by MAP estimation [26]. Although any components (means, covariances and weights) can be adapted, mean adaptation is commonly adopted and this approach is used in our study. With this setup, the likelihood of a test utterance $x = \{x_t; t = 1, 2, \ldots, T\}$ evaluated on the model of a speaker s is given by:

$$L(x; s) = \prod_t \sum_k \pi_k \mathcal{N}(x_t; \mu_k^s, \Sigma_k) \tag{1}$$

where x_t is the speech feature vector at frame t, and k indexes the Gaussian component.$\mathcal{N}(\cdot; \mu_k^s, \Sigma_k)$ is the k-th Gaussian component with the adapted mean vector μ_k^s and the covariance matrix Σ_k, and π_k is the associated prior distribution. We highlight that here $\{\pi_k\}$ are speaker independent since they are not updated in speaker enrollment. This means that if the true distribution of an enrollment speech deviates from the model prior, the enrolled model will be biased. Likewise, if the true distribution of a test speech deviates from the prior, the likelihood score for the test speech will be biased.

If the enrollment/test speech is sufficiently long, the true distribution of the speech tends to match the model prior well, partly due to the fact that speech signals of a particular language follow a certain natural distribution over phones. However, if the enrollment/test speech is short, the model prior usually can not reflect the true distribution of the signal, leading to biased speaker models and biased likelihood evaluation.

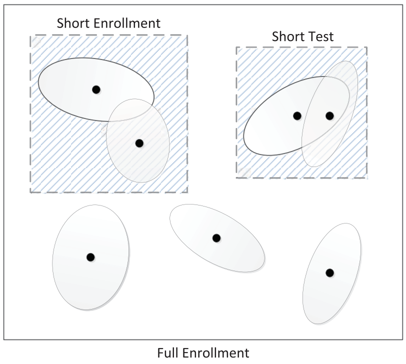

Fig. 1. Mismatch between the model prior and the true distributions of enrollment/test speech signals.

The problem of prior-mismatch is show in Fig. 1, where the ellipses represent Gaussian components, and the two squares represent the coverage of the enrollment and test speech respectively. If the enrollment speech is sufficient, there is less problem with the prior mismatch issue and the speaker model can be well trained (the outer large square); however since the test speech is short and so only part of the Gaussian components are covered, the likelihood evaluation is biased. This is reflected by the fact that computing the likelihood is impacted by the Gaussian components that are not covered by the test speech. If the enrollment utterance is short as well, the components covered by the enrollment and test speech could be even non-overlapping. This causes more severe problem because: (1) the components covered by the test speech are not well trained in enrollment; (2) the components that are trained in enrollment are not the ones covered (required) by the test speech, which in turn impacts the accuracy of the likelihood estimate.

This paper proposes a subregion modeling approach to tackle this problem. Specifically, the acoustic feature space is divided into a number of 'homogeneous' subregions, where 'homogeneous' means that the above mentioned matched-priori assumption is satisfied. The UBM and speaker GMMs are then constructed within each subregion, and the likelihood is computed by merging the evaluations on all the individual subregion models. This can be formulated as follows:

$$L(x; s) = \prod_t \sum_c P(c|x_t) \sum_k \pi_{c,k} \mathcal{N}(x_t; \mu_{c,k}^s, \Sigma_{c,k}) \tag{2}$$

where c indexes the regions, and $P(c|x_t)$ is the posterior probability that x_t resides in the c-th subregion. This model can be simplified by a 'hard' subregion assignment, given by:

$$L(x; s) \approx \prod_t \sum_k \pi_{\tilde{c},k} \mathcal{N}(x_t; \mu_{\tilde{c},k}^s, \Sigma_{\tilde{c},k}) \tag{3}$$

where $\tilde{c}$ denotes the subregion that is assigned to x_t by MAP, given by:

$$\tilde{c} = \arg\max_c P(c|x_t).$$

The central task of the above subregion modeling is to define the subregions and estimate the posterior probability $P(c|x_t)$.

This can be achieved by clustering the Gaussian components in an unsupervised fashion and then computing $P(c|x_t)$ by the Bayesian rule, but this is usually not satisfactory as the unsupervised learning does not leverage any external knowledge so the resulting model would not be very different from a larger GMM with more Gaussian components. A more ideal approach is to associate each subregion c with a speech unit, e.g., a phone. We choose this approach and employ an automatic speech recognition (ASR) system to conduct the subregion assigned by the technique of forced phone alignment. This approach possesses several advantage. First, it is a supervised clustering that involves linguistic knowledge, e.g., the phone inventory, and so the constructed subregions tend to be homogeneous in nature. Second, by employing ASR, it implicitly leverages much exotic resources that are used to train the ASR system, e.g., large speech data, word dictionaries and language models. Third, with the phones obtained with ASR, it is possible to choose the best discriminative subregions, such as those associated with vowels or nasals.

With the subregion modeling, speakers can be modeled in a more thorough way, given that sufficient training data are available for each speech unit. In practice, however, data are often scarce for some speech units. This paper proposes a solution which clusters similar speech units into speech unit classes, and uses the speech unit classes to construct robust acoustic subregions. This approach works well with sufficient enrollment data as we will show in Section V; however, if the enrollment utterance is short, it is still problematic. This is because some speech unit classes may be assigned very little or even no enrollment data, and so the corresponding subregion speaker models are highly under-estimated. To solve this problem, a model synthesis approach is proposed in this paper, which synthesizes models for speech unit classes with very little training data from classes with abundant data by a linear transform.

The rest of the paper is organized as follows: Section II discusses some related works, and III presents the subregion modeling, where we assume that the enrollment data is sufficient. Section IV presents the model synthesis approach to deal with speech units with limited enrollment data. Section V describes the experiments, and the entire paper is concluded in Section VI.

II. Related Work

The idea of employing phonetic information in speaker recognition has been investigated by previous research studies. For instance, Omar et al. [27] proposed to derive UBMs from Gaussian components of a GMM-based ASR system, with a K-means clustering approach based on symmetric KL distance. Another work is the DNN-based i-vector method proposed by Lei and colleagues [16]. In their work, posteriors of senones (context-dependent states) generated by a DNN trained for ASR were used for model training as well as i-vector inference.

The subregion model presented here follows the same idea of exploiting phonetic knowledge learned by ASR systems. The difference is how the knowledge is used. Omar's work uses the GMM-based acoustic model to construct robust UBMs, and Lei's work uses DNN-based acoustic model to generate component posteriors for model training and inference. In contrast, the subregion approach presented here uses a full-fledged ASR system to generate phone labels. An advantage of our approach is that strong language models can be applied to offer more accurate phone labels; additionally, the acoustic classes (subregions) are modeled explicitly in our proposal, which is highly flexible. For example, although GMMs are used in the present study, it can be any generative model such as the i-vector model.

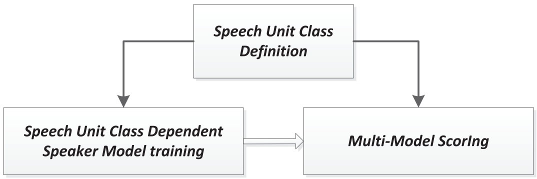
Speech Unit Class Definition
Speech Unit Class Dependent Speaker Model training
Multi-Model Scorlng

Fig. 2. The Framework of the subregion modeling.

Regarding the research for SUSR, it has been known that the i-vector model possesses some advantages when dealing with short enrollment/test utterances [23]. This may be largely attributed to the nature of this model in sharing statistical strength among different acoustic regions. The subregion model tackles the SUSR problem in a different way: it relies on conditional models that describe speech signals in the most appropriate acoustic classes. We believe that these two approaches can be combined in a certain way but leave the investigation as future work.

III. Subregion Modeling Based on Speech Unit Classes

The proposed subregion framework involves three components. Firstly the speech unit classes are derived by clustering similar speech units. Secondly the subregion models (including UBMs and speaker GMMs) are trained for each subregion that is defined by the speech unit classes. Finally test utterances are scored with the subregion models. Fig. 2 illustrates the system framework.

A. Speech Units Based on Finals

The inventory of speech units varies for different languages. In Chinese, the language focused in this paper, speech units can be words, syllables, Initials/Finals (IF) or phones [28]. Although language-independent speech units can be defined, e.g., through the International Phonetic Association (IPA) [29] and multi-lingual speaker/speech recognition systems [30], [31], language-dependent speech units generally cover the acoustic space in a better way. Therefore we consider Chinese-specific speech units to define the subregions in this paper.

A widely used speech unit definition in Chinese is based on the Initial/Final (IF) structure of syllables, where the initials correspond to consonants, and the finals correspond to vowels and nasals [28]. Compared with other units such as syllables and phones, the IFs are moderate in number (65 in total) and can reflect the phonetic structure of Chinese pronunciations. The IF set has been reproduced in Table I, where {_a, _o, _e, _i, _u, _v} are zero initials and appear in non-initial syllables [28].

TABLE I
THE IF SET OF STANDARD CHINESE

Type	Units
Initial (27)	b, p, m, f, d, t, n, l, g, k, h, j, q, x, zh, ch, sh, z, c, s, r, _a, _o, _e, _i, _u, _v
Final (38)	a, ai, an, ang, ao, e, ei, en, eng, er, o, ong, ou, i, i1, i2, ia, ian, iang, iao, ie, in, ing, iong, iou, u, ua, uai, uan, uang, uei, uen, ueng, uo, v, van, ve, vn

Among the IFs, Finals have been found to convey more speaker dependent/specific information than Initials [32], [33]. Better speaker recognition performance therefore can be obtained by selecting speech segments corresponding to Finals only. In other words, the Finals are the effective speech units when constructing subregion models in this study.

B. Speech Units Clustering

Once the speech units are defined as the Finals, the subregion modeling can be conducted by building Final-dependent GMM-UBMs. This approach, however, is almost impossible in practice, due to data sparsity caused by the large number of Finals. A possible solution is to cluster similar units together and build subregion models based on the resulting speech unit classes. Two clustering approaches are investigated in this section, one is based on phonetic knowledge and the other is data-driven.

1) Clustering by Phonetic Knowledge: The first approach clusters the Finals based on phonetic knowledge. This paper directly applies the definition of speech unit classes provided by [34], which is based on tongue's height and backness information of the speech units in the IPA definition.

2) Clustering in Data-Driven Way: The second approach clusters the Finals based on the distributions of speech signals of each Final. There are a multitude of approaches to this clustering, e.g., the tree-based tying used for acoustic modeling in ASR [35] and unit selection in speech synthesis [36], the greedy merge of similar classes used in maximum likelihood linear regression (MLLR) [37], [38]. Most of these approaches try various possible merge schemes and select the best one that leads to maximum likelihood on training data. In this study, we develop a vector quantization (VQ) method based on the K-means algorithm [39] to conduct the clustering. In contrast to the methods mentioned above, our approach calculates pair-wised distance among models, and then select close models to merge. Since no training data need to be revisited for every possible clustering schemes, our method is simple and quick. A regression tree-based method which utilizes both data and knowledge of phonetic classes tends to get better clusters. However, since the clustering method itself is not the main focus of this work, the simple K-means algorithm was used in this study. Note that a similar approach has been employed in [27].

The whole clustering process is illustrated as follows:

- Train a global UBM with a large training dataset. The data are chosen to cover all the Finals, and are balanced in terms of genders.
- Let N denote the number of Finals. Collect data of each Final and train local (Final-dependent) UBMs based on the global UBM by MAP. Again, the off-the-shelf speech recognition system is employed to segment the training speech data. Denote the local UBM of Final i by $\lambda_i = \{\pi_k, \mu_{i,k}, \Sigma_k : k = 1, \ldots, K\}$. Note that only $\{\mu_{i,k} : k = 1, \ldots, K\}$ are Final-dependent.
- Define the distance of two Final-dependent UBMs based on the symmetric Kullback-Leibler (KL) divergence [40], given by:

$$\lambda_i \| \lambda_j = \sum_{k=1}^{K} \pi_k (N(\mu_{i,k}, \Sigma_k) \| N(\mu_{j,k}, \Sigma_k)) \quad (4)$$

where

$$N(\mu_{i,k}, \Sigma_k) \| N(\mu_{j,k}, \Sigma_k) = \sum_{d=1}^{D} \frac{(\mu_{i,k}(d) - \mu_{j,k}(d))^2}{\sigma_k(d)^2},$$

where D is the dimension of the feature vector. Note we have assumed that the covariance matrices are diagonal, and the d-th primary diagonal element has been denoted by $\sigma(d)$.
- Assume that the number of unit clusters requested is C. Select C Final-dependent UBMs as the initial centers of the C classes. The selection is based on the KL divergence defined above and applies the max-min criterion, i.e., sequentially select the UBM whose minimum distance to other UBMs is the maximum.
- The K-means algorithm [39] is conducted to cluster the N Final-dependent UBMs into C clusters, with the distance measure set to the KL divergence.

C. Subregion Modeling Based on Speech Unit Classes

Denote the speech unit classes (Final clusters) by $\{\text{SUC-}c := 1, \ldots, C\}$. Based on the classes, a subregion UBM can be trained for each SUC-c with the training data that are aligned to the Finals in SUC-c by the speech recognition system. The subregion UBM of class SUC-c is denoted by λ_c^{UBM}. The speaker-dependent subregion GMM models can be trained based on the subregion UBMs, using the enrollment data that have been aligned to the Finals.

In summary, the entire process of the subregion modeling approach is illustrated in Fig. 3, and the details are as follows:

- Global UBM training, denoted by λ^{UBM}. A global UBM is trained with the entire training dataset by employing the expectation-maximization (EM) algorithm [25], [41].
- Subregion UBM training. The speech recognition system is used to align the speech signals (acoustic features) to the Finals. The aligned speech data are then assigned to the C speech unit classes according to the definition of $\{\text{SUC-}c\}$. A subregion UBM λ_c^{UBM} is trained for the c-th speech unit class based on the global UBM, by employing the MAP algorithm [26] and with the speech data assigned to SUC-c.

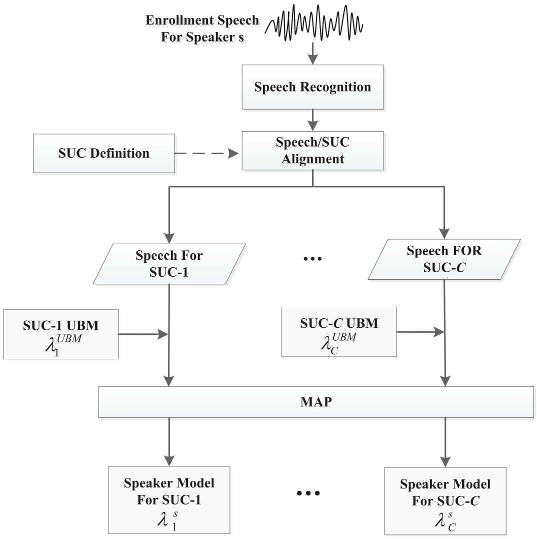

Fig. 3. Speaker-dependent subregion model training. 'SUC' stands for speech unit class.

- Subregion speaker model training. For a speaker s, first segment the enrollment speech data into Finals and assign the speech data to the speech unit classes, by the same way as in the subregion UBM training. Then for each speech unit class SUC-c, a subregion speaker-dependent GMM λ_c^s is trained by MAP adaption from the subregion UBM λ_c^{UBM} with the assigned enrollment data.

Note that with the subregion model, the total parameters of a speaker model would be significantly increased, possibly leading to the problem of data sparsity. However, the problem is not as that serious as the first glance, because only the mean vectors are updated while priors and variances are shared across subregions. Nevertheless, it would be certainly good if some pruning approach is applied to remove unrepresentative Gaussian components. We leave this pruning method as future work.

D. Scoring With Subregion Models

With the speaker-dependent subregion GMMs trained, a test utterance can be scored by scoring on each subregion and taking the average. More sophisticated approach to fuse the subregion scores is left for future study. Suppose a test utterance contains L Finals according to the decoding result of speech recognition, and denotes the speech unit class of the l-th final by $c(l)$. Further denote the speech segment of this unit by X_l, and its length is T_l. The score of X_l is measured by the log likelihood ratio between the subregion speaker-dependent GMM $\lambda_{c(l)}^s$ and the subregion UBM $\lambda_{c(l)}^{UBM}$, where s denotes the speaker. This is formulated by:

$$\varphi_{i,l} = \log p(\mathbf{X}_l | \lambda_{c(l)}^i) - \log p(\mathbf{X}_l | \lambda_{c(l)}^{ubm})$$

The score of the entire utterance is computed as the average of the segment-based scores:

$$\varphi_i = \frac{\sum_{l=1}^L \varphi_{i,l}}{\sum_{l=1}^L T_l}.$$

IV. Speaker Model Synthesis

The subregion modeling presented in the previous section models and scores speech signals in appropriate subregions, and therefore does not rely on the global prior distribution, i.e., $\{\pi_k\}$ in (1). If all the subregion models are well trained, then a major difficulty associated with SUSR, i.e., the biased prior distribution caused by short *test* utterances, is largely solved.

A potential problem of this approach is that if the *enrollment* utterance is short as well, some of the subregion models can be under-estimated, which will lead to significant performance reduction if the test utterances fall in the data-sparse subregions. The unit clustering approach discussed in the previous section can partially solve the problem, however it is still problematic if the enrollment utterance is very short. In this section, we propose a model synthesis approach to address the problem, which constructs subregion models for speech unit classes with no or very limited enrollment data based on data-rich subregion models by a linear transform. The basic assumption is that the relationship between two subregion models does not change when speaker-dependent models (subregion GMMs) are adapted from speaker-independent models (subregion UBMs), and the relationship can be represented by a linear transform. These transforms can then be applied to synthesize speaker-dependent GMMs for speech unit classes with limited data. In this study, we employ the maximum likelihood linear regression to train the linear transform.

A. Maximum Likelihood Linear Regression

The maximum likelihood linear regression (MLLR) [37], [42] was first proposed by the Cambridge group to deal with channel mismatch and speaker variability in speech recognition. Given a GMM $\lambda = (\pi_k, \mu_k, \Sigma_k : k = 1, 2, \ldots, K)$ and a speech segment X, the MLLR seeks a linear transform L that maximizes the likelihood function

$$P(X; \lambda, L) = \sum_k \pi_k N(X; L\xi_k, \Sigma_k) \tag{5}$$

where

$$\xi_k = [\mu_{k,1}, \ldots, \mu_{k,D}, 1]$$

is the extended mean vector, and D is the dimension of speech features. L is an $D \times (D+1)$ transformation matrix. The optimization of the matrix L in the sense of maximum likelihood gives the following estimation:

$$L_i = \kappa_i {G_i}^{-1}$$

where L_i is the i-th row of L, and κ_i, ${G_i}^{-1}$ are calculated as:

$$\kappa_i = \sum_{k=1}^K \sum_{t=1}^T r_k(t) \frac{1}{{\sigma_k}^2(i)} x_i(t) {\xi_k}^T$$

$$G_i = \sum_{k=1}^K \frac{1}{{\sigma_k}^2(i)} \xi_k {\xi_k}^T \sum_{t=1}^T r_k(t)$$

where t indexes time, $x_i(t)$ is the i-th element of the feature vector at time t, and $r_k(t)$ is the posterior probability of $x(t)$

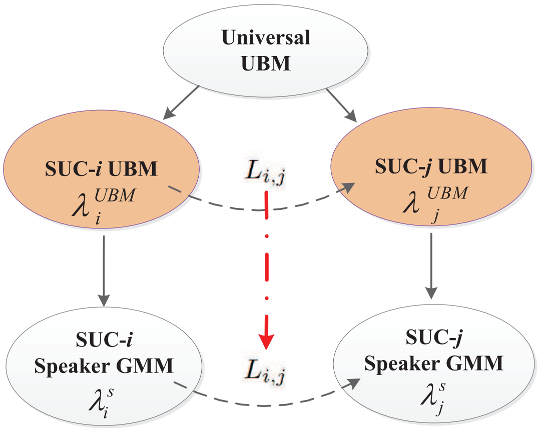

Fig. 4. Illustration of model synthesis based on subregion UBMs. 'SUC' stands for speech unit class.

which belongs to the k-th Gaussian component. ${\sigma_k}^2(i)$ is the i-th primary diagonal element of Σ_k, where we have assumed that Σ_k is diagonal.

B. Model Synthesis Based on Subregion UBMs

With the MLLR technique, a transform $L_{i,j}$ can be learned for each subregion UBM pair (λ_i^{UBM}, λ_j^{UBM}). Since the amount of speech data aligned to each speech unit class is relatively large when training the subregion UBMs, the transforms can be easily learned. For example, to learn $L_{i,j}$, the subregion UBM λ_i^{UBM} is used as the GMM model in (5), and the speech data aligned to the j-th speech unit class are used as the adaption data X.

Once the transforms are learned, they can be used to synthesize speaker-dependent subregion GMMs in speaker enrollment. Specifically, the enrollment speech data is first segmented by the speech recognition system and the speech features are assigned to the speech unit classes. If a speech unit class j involves sufficient training data, then the subregion GMM λ_j^s is derived by MAP from the corresponding subregion UBM λ_j^{UBM}, where s denotes the speaker. If the speech unit class involves little training data, then the subregion GMM is synthesized from a well-trained speaker-dependent subregion model, λ_i^s for example. The synthesis is implemented as a linear transform:

$$\mu_{j,k} = L_{i,j} \begin{bmatrix} \mu_{i,k} \\ 1 \end{bmatrix} \quad k = 1, 2, \ldots, K$$

where k indexes the Gaussian components.

Fig. 4 illustrates the subregion UBM-based model synthesis. Firstly the transform $L_{i,j}$ is learned to map the subregion UBM λ_i^{UBM} to λ_j^{UBM}, and then $L_{i,j}$ is used to synthesize the speaker subregion GMM λ_j^s based on λ_i^s.

C. Model Synthesis Based on Cohort Speakers

A particular shortcoming of the subregion UBM-based model synthesis is that the transforms $\{L_{i,j}\}$ are speaker independent. This is a strong assumption, as different speakers may exhibit completely different characteristics when moving from one pronunciation to another. We propose the speaker-dependent transforms based on cohort sets.

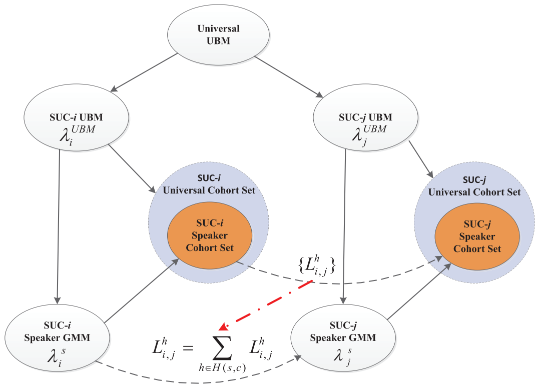

Fig. 5. The illustration of model synthesis based on cohort speakers. 'SUC' stands for speech unit class.

A cohort set [43] is a cluster of speakers that share similar characteristics. Given a speaker s, there is an individual cohort set $H(s, c)$ for each subregion c, and every cohort set $H(s, c)$ involves speakers that are similar to speaker s in the c-th subregion. The KL divergence is used to measure speaker distance in our study, as given by (4).

The cohort speaker-based synthesis is illustrated in Fig. 5. Firstly we chose a universal cohort speaker set H which involved 300 speakers, and each speaker was modeled by a set of subregion GMMs, defined as $\{\lambda_c^h : c = 1, 2, \ldots, C\}$, where h indexes the speaker and c indexes the subregion. Secondly the MLLR transform was estimated for each speaker h between each subregion pair (i, j), denoted by $\{L_{i,j}^h : h \in H\}$.

When registering a speaker s, for each speech unit class c, if the training data are sufficient, the subregion speaker model λ_c^s is trained directly by MAP with the corresponding subregion UBM λ_c^{UBM}; otherwise, it is synthesized from subregion models of his/her cohort speakers. Specifically, specify a data-rich subregion of the speaker, e.g., subregion c', and then specify the cohort set $H(s, c') \subset H$ by finding the similar speakers in the universal cohort set H. The subregion model λ_c^s for data-sparse subregion c is then synthesized from the data-rich subregion model of speaker s, i.e., $\lambda_{c'}^s$ and the linear transforms defined by the cohort set, that is $\{L_{c',c}^h : h \in H(s, c')\}$. Again, only the mean vectors are synthesized, formulated by:

$$\mu_{c,k}^s = \sum_{h \in H(s,c')} L_{c',c}^h \begin{bmatrix} \mu_{c',k}^s \\ 1 \end{bmatrix} \quad k = 1, 2, \ldots, K$$

where k indexes the Gaussian components.

V. Experiments

A. Database

1) Database for Evaluation (SUD12): There is not a standard database for performance evaluation on text-independent SUSR tasks. A possible way to construct an SUSR database quickly is to cutting out words or phrases from a database used for general speaker recognition. This approach, however,

TABLE II
DI-IF STATISTICS OF SUD12 ENROLLMENT DATA

di-IF Type	Example	Number
Initial - Final	zh-ong	380
Zero Initial - Final	_y-uan	36
Final - Initial	ong-n	798
Final - Zero Initial	ua-_y	228
All	–	1,442

TABLE III
LENGTH DISTRIBUTION OF SUD12 TEST DATA

Length (s)	Number	Percentage (%)
≤ 0.5	38	60.3
0.5 - 1.0	15	23.8
1.0 - 2.0	10	15.9

may introduce artifacts when cutting continuous speech signals. We therefore decided to design and record a database that is suitable for SUSR research and publish it for research usage[1]. The database was named as "SUD12" [44], [45], and was designed in the principle to guarantee sufficient IF balance. In order to focus on short utterances and exclude other factors such as channel and emotion, the recording was conducted in the same room and with the same microphone, and the reading style was neutral. There are in total 28 male speakers and 28 female speakers, and all the utterances are in Standard Chinese. The sampling rate is 16 kHz, and the sampling precision is 16 bits. For each speaker, there are 100 Chinese sentences, each of which contains 15 $\sim$ 30 Chinese characters. These sentences were selected by the ELFU algorithm [46] from $5,000$ sentences in the news domain taken from the Peoples Daily, with the objective to maximize the di-IF coverage [47]. The IF coverage rate is 100% and the di-IF coverage rate is 82%, and each IF exists in at least 10 utterances. The statistics of the di-IF is presented in Table II.

The enrollment dataset involves all the 56 speakers. For each speaker, 10 utterances are randomly selected and merged together as the enrollment speech. After removing the silence segments, the effective speech signals for enrollment is about 35 seconds.

The test dataset of SUD12 involves 56 speakers, and each speaker speaks 62–63 short utterances, which covered all the Finals in Standard Chinese. The lengths of the recordings are not more than 2 seconds and mostly shorter than 0.5 seconds. The distribution is shown in Table III. The evaluation involves $3,523$ target trials and $197,293$ non-target trials.

2) Database for UBM Training (863DB): The speech data used to train the UBMs and subregion UBMs were chosen from the 863 Chinese speech corpus [48]. The 863 database was well designed to cover all the Chinese IFs, and which is particularly suitable to train subregion UBMs for speech unit classes. All the recordings are at a sampling rate of 16 kHz, and the sample precision is 16 bits. In this study, we chose 38 males and 33 females from the 863 corpus, and for each speaker, there are 150 speech utterances in average, and the length of the speech signals is 17 hours in total. This dataset is denoted by 863DB for convenience.

3) Database for Cohort Speaker Selection (dEarDB): In order to construct cohort-based MLLR transforms, we employed another cohort speaker database that was recorded by Beijing d-Ear Technologies Co., Ltd. for Korea Speech Information Technology and Promotion Center. It contains 150 male speakers and 150 female speakers. As SUD12, the recordings are sampled at 16 kHz with 16-bit precision. For each speaker, 250 Standard Chinese sentences were recorded, and the effective speech content of each utterance is approximately 3 seconds long. This database is denoted by dEarDB.

B. Experimental Conditions

The Kaldi toolkit [49] was adopted to conduct the experiments, and the recipe to reproduce the results can be found online[2]. Following the standard recipe of SRE08, the acoustic feature is the conventional 60-dimensional Mel frequency cepstral coefficients (MFCC), which involves 20-dimensional static components plus the first and second order derivatives. The frame size is 25 ms and the frame shift is 10 ms. The number of mel-frequency bins is 23 and the frequency range is from 20 Hz to $8,000$ Hz.

Note that a simple energy-based voice activity detection (VAD) is performed before the feature extraction, and the cepstral mean normalization (CMN) [50] is applied as a post-process to reduce the impact of channel mismatch.

We chose the conventional GMM-UBM approach to construct the baseline system. The UBM consisted of $1,024$ Gaussian components and was trained with the 863DB. Note that this setting is 'almost' optimal in our experiments, i.e., using more Gaussian components can not improve system performance in any significant way. The SUD12 was employed to conduct the evaluation. With the enrollment data, the speaker GMMs were derived from the UBM by MAP, where the MAP adaptation factor was optimized so that the EER on the test set was the best. The final result on the SUD12 test set is 28.97% in EER. This is a reasonable performance for SUSR that involves short utterances less than 2 seconds [21], [22].

The ASR system used to generate the phone alignment was a large scale DNN-HMM hybrid system. The system was trained using Kaldi following the WSJ S5 recipe. The feature used is 40-dimensional Fbanks. The basic features are spliced by a window of 11 frames, and an LDA (linear discriminative analysis) transform is applied to reduce the dimension to 200. The DNN structure involves 4 hidden layers, each containing $1,200$ hidden units. The output layer contains $6,761$ units, corresponding to the number of GMM senones. The DNN was trained with $6,000$ hours of speech signals, and the decoding employed a powerful 5-gram language model trained on 2 TB text data.

For comparison, a GMM-based i-vector system and DNN-based i-vector system were also constructed. The GMM-based i-vector system used the same UBM model as the GMM-UBM system, and the dimension of the i-vector is 400. For the DNN-based i-vector system, the DNN model was trained following

[1] http://www.cslt.org/resources.php?Public%20data

[2] http://lilt.cslt.org

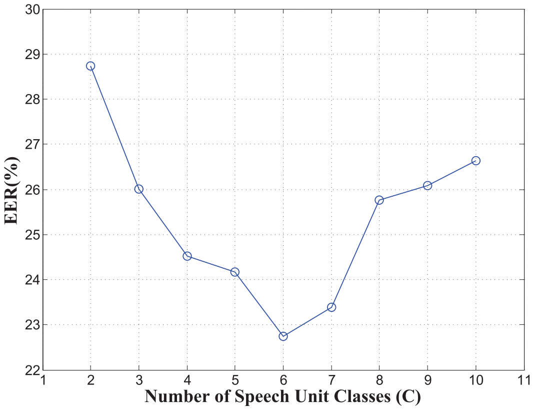

Fig. 6. EERs with different numbers of speech unit classes in data-driven clustering.

TABLE IV
SPEECH UNIT CLASSES DERIVED IN DATA-DRIVEN WAY

Class	Speech Units
1	a, ao, an, ang, ia, iao, ua
2	e, ei, ai, i, ie, uei, iii
3	iou, ou, u, ong, uo, o
4	v, vn, ve, van, er
5	en, ian, uan, uen, uai, in, ii, ing
6	eng, iang, iong, uang, ueng

the same procedure as the one used for the ASR system, but with less number of senones. In our experiments, the number is 928, comparable to the number of Gaussian components of the GMM-UBM system.[3] The dimension of the i-vector space is set to 400, and the posteriors produced by the DNN model are used in both model training and i-vector extraction.

C. Subregion Modeling

The first experiment investigates the subregion modeling based on speech unit clustering. Two clustering approaches are studied: the knowledge-based approach (‘SBM-KW’) and the data-driven approach (‘SBM-DD’). For the knowledge-based approach, we simply follow the definition of speech unit classes described in [34]. For the data-driven approach, it is necessary to choose an appropriate number of classes for the clustering algorithm. If the number of classes is small, the subregions tend to be not homogeneous in terms of prior distributions and so can not well deal with short test utterances, and if the number of classes is large, the problem of data sparsity is more serious. In order to determine the optimal class number (denoted by C), the recognition performance with various values of C has been evaluated and the results are reported in Fig. 6. It can be seen that either too small or too large values lead to suboptimal performance, and the optimal setting in our experiment is C=6. Table IV shows the derived unit classes with this configure. It can be seen that the clustering result is reasonable at least intuitively.

TABLE V
PERFORMANCE COMPARISON

System	EER (%)
GMM-UBM (baseline)	28.97
GMM i-vector	39.91
DNN i-vector	29.61
SBM-KW	24.30
SBM-DD	22.74

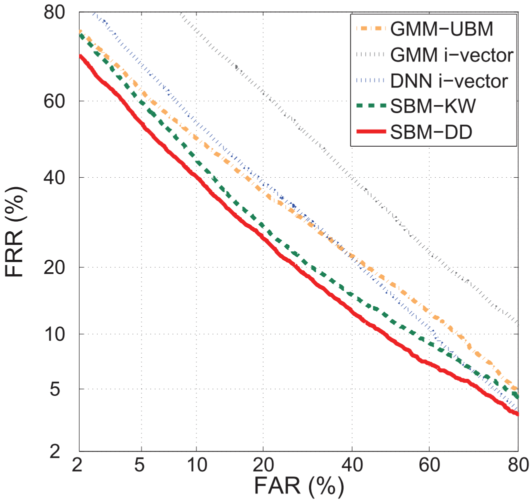

Fig. 7. The DET curves with the GMM-UBM/i-vector baselines, and two subregion modeling methods.

The results in terms of EER are presented in Table V, where ‘GMM-UBM’ is the GMM-UBM baseline system, ‘GMM i-vector’ denotes the traditional GMM-based i-vector system, and ‘DNN i-vector’ denotes the DNN-based i-vector system.

‘SBM-KW’ and ‘SBM-DD’ are subregion systems with the knowledge-based and data-driven speech unit clustering, respectively. Note that the optimal number of classes (C=6) has been employed in the data-driven system. For a better understanding of the performance on various operation points, the DET curves are presented in Fig. 7, where the horizontal axis represents the false acceptance rate (FAR) and the vertical axis represents the false rejection rate (FRR) [51].

We first observe that the GMM-UBM baseline outperforms the two i-vector systems. This might be largely because no normalization methods (e.g., PLDA) are applied to the i-vector systems. The DNN-based i-vector system outperforms the GMM-based i-vector system in a significant way. This is also expected as the phonetic knowledge was employed by the DNN-based system.

Furthermore, it can be seen that the systems based on subregion modeling outperform the GMM-UBM baseline, with either the knowledge-based or data-driven clustering approach. When comparing the two clustering approaches, it is observed that the data-driven approach is more effective. This is probably because the data-driven approach takes into account the characteristics of real data, and the balance of data over the resultant speech unit classes may have lead to more robust subregion models.

[3] Note that it is not easy to set the exact number of senones in the ASR system with the tree-based clustering algorithm for context-dependent states.

TABLE VI
RESULTS WITH MODEL SYNTHESIS BASED ON SUBREGION UBMS

EER (%)	SBB1	SBB2	SBB3	SBB4	SBB5	SBB6	Average	NULL
SBS1	–	23.84	23.90	25.06	23.96	24.87	24.33	35.59
SBS2	25.57	–	25.15	26.51	24.98	26.20	25.68	33.30
SBS3	25.66	25.23	–	27.85	27.25	26.74	26.55	32.76
SBS4	22.74	22.08	22.79	–	22.11	23.08	22.56	26.74
SBS5	25.75	23.84	25.86	26.77	–	25.38	25.52	26.40
SBS6	23.67	23.59	23.56	23.99	24.81	–	23.92	23.64

One may argue that the comparison in Table V is not completely fair, as the subregion model involves more parameters and thus naturally more powerful. This is certainly true in general, however in practical systems where training and enrollment data are limited, more complex models unnecessarily deliver better performance. In fact in our experiment, it showed that 1024 Gaussian components are sufficient for the conventional GMM-UBM model to describe the entire acoustic space (at least with the current modeling approach based on EM/MAP) and adding more components did not offer clear advantage. Therefore, the gains obtained by the subregion modeling should not be attributed to the increased parameters, but the new modeling method based on subregions that are derived from the external speech recognition system.

D. Model Synthesis

The second experiment studies the MLLR-based model synthesis for speech unit classes with very little enrollment data. We choose the class definition in Table IV, and simulate data-sparse speech unit classes by discarding the speech segments assigned to the class.

1) Synthesis Based on Subregion UBMs: We study the model synthesis approach based on subregion UBMs. The results are shown in Table VI, where the value shown in the element (SBSi,SBBj) is the EER with the i-th subregion model synthesized from the j-th subregion model. The column 'Average' presents the averaged EER over all the subregion j. The column 'NULL' presents the results without any model synthesis and here it is regarded as the baseline system. It can be seen that with the model synthesis, the performance is generally improved compared with the baseline system. An exception is the subregion 6, for which the synthesis does not work well because the pronunciations in this acoustic class is absent. Checking Table IV, one can find that most of the phones in this class are ended with the nasal 'ng'. It seems to indicate that nasal-ended Finals are difficult to be synthesized. Moreover, the pronunciations of this class take only a small proportion of the entire test dataset, and therefore the result presented here is not statistically significant.

2) Synthesis Based on Cohort Speakers: As discussed in Section IV, synthesis based on subregion UBMs suffers from the speaker-independent assumption for MLLR transforms. This experiment studies the speaker-dependent synthesis approach based on speaker-dependent cohort sets. For simplicity, we choose the 3-th speech unit class as the data-sparse class and synthesize the subregion model from the model of the 4-th speech unit class.

TABLE VII
RESULTS WITH MODEL SYNTHESIS

System	EER (%)
SBM-DD	22.74
NO-MLLR (baseline)	32.76
MLLR-UBM	27.85
MLLR-COHORT	27.53

Firstly we investigate the impact of the size of the speaker-dependent cohort set. It was found that the EER first drops as the size of the speaker-dependent cohort set increases, until the best performance is reached; afterward, the EER starts to increase as the size of the cohort set increases. In our experiment, the best result is obtained when the size of the cohort set is set to 80. This optimal value is used in the rest of the experiments.

Table VII presents the results with the MLLR-based model synthesis, where the row 'NO-MLLR' presents the system without any treatment for the data-sparse speech unit class. Compared with the case with sufficient enrollment data ('SMB-DD'), significant performance reduction is observed. This means that enrollment data sparsity indeed causes serious impact for speaker recognition. The row 'MLLR-UBM' presents the system with model synthesis based on subregion UBMs, and the row 'MLLR-COHORT' presents the system with model synthesis based on speaker-dependent cohort sets. It can be found that model synthesis does offer clear performance improvement in the case with limited enrollment data, and the cohort-set-based synthesis slightly outperforms the subregion UBM-based synthesis.

VI. CONCLUSION AND FUTURE WORK

In this paper, we propose a subregion modeling approach for text-independent short utterance speaker recognition. To deal with the problem of data sparsity in enrollment and test, the speech units (IFs) are clustered into speech unit classes in the subregion modeling; and to deal with short enrollment utterances, a model synthesis approach based on MLLR has been proposed. The experimental results show that the proposed subregion modeling approach, plus the data-driven speech unit clustering, gains significant performance improvement on very short test utterances. In the case of limited enrollment data, the simulation experiment shows that the model synthesis approach based on both the subregion UBMs and cohort speakers can largely recover the performance lost caused by enrollment data sparsity. Future work involves combination of feature-based and model-based compensations for short utterances, and testing the proposed approaches in the i-vector framework.

REFERENCES

[1] J. P. Campbell, Jr., “Speaker recognition: A tutorial,” *Proc. IEEE*, vol. 85, no. 9, pp. 1437–1462, Sep. 1997.

[2] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo, “Support vector machines for speaker and language recognition,” *Comput. Speech Lang.*, vol. 20, no. 2, pp. 210–229, 2006.

[3] F. Bimbot *et al.*, “A tutorial on text-independent speaker verification,” *EURASIP J. Appl. Signal Process.*, vol. 2004, pp. 430–451, 2004.

[4] T. Kinnunen and H. Li, “An overview of text-independent speaker recognition: From features to supervectors,” *Speech Commun.*, vol. 52, no. 1, pp. 12–40, 2010.

[5] C. S. Greenberg *et al.*, “The 2012 NIST speaker recognition evaluation,” in *Proc. INTERSPEECH*, 2013, pp. 1971–1975.

[6] D. A. Reynolds, “Speaker identification and verification using gaussian mixture speaker models,” *Speech Commun.*, vol. 17, no. 1, pp. 91–108, 1995.

[7] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted Gaussian mixture models,” *Digit. Signal Process.*, vol. 10, nos. 1–3, pp. 19–41, 2000.

[8] J.-L. Gauvain and C.-H. Lee, “Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains,” *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

[9] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, “Joint factor analysis versus eigenchannels in speaker recognition,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1435–1447, May 2007.

[10] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 788–798, May 2011.

[11] A. O. Hatch, S. S. Kajarekar, and A. Stolcke, “Within-class covariance normalization for SVM-based speaker recognition,” in *Proc. INTERSPEECH*, 2006, pp. 1471–1474.

[12] M. Senoussaoui, P. Kenny, N. Dehak, and P. Dumouchel, “An i-vector extractor suitable for speaker recognition with both microphone and telephone speech,” in *Proc. Odyssey*, 2010, pp. 28–33.

[13] S. J. Prince and J. H. Elder, “Probabilistic linear discriminant analysis for inferences about identity,” in *Proc. IEEE 1th Int. Conf. Comput. Vis. (ICCV’07)*, 2007, pp. 1–8.

[14] L. Deng and D. Yu, “Deep Learning: Methods and Applications,” *J. Found. Trends Signal Process.*, vol. 7, no. 3–4, pp. 197–387, Jun. 2014, Hanover, MA, USA: Now.

[15] P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet, and J. Alam, “Deep neural networks for extracting baum-welch statistics for speaker recognition,” in *Proc. Odyseey*, 2014, pp. 293–298.

[16] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, “A novel scheme for speaker recognition using a phonetically-aware deep neural network,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2014, pp. 1695–1699.

[17] V. Ehsan, L. Xin, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, “Deep neural networks for small footprint text-dependent speaker verification,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2014, pp. 4052–4056.

[18] L. Li, D. Wang, Z. Zhang, and T. F. Zheng, “Deep speaker vectors for semi text-independent speaker verification,” arXiv preprint arXiv:1505.06427, 2015.

[19] A. Larcher, K.-A. Lee, B. Ma, and H. Li, “RSR2015: Database for text-dependent speaker verification using multiple pass-phrases,” in *Proc. INTERSPEECH*, 2012, pp. 1580–1583.

[20] R. Vogt, S. Sridharan, and M. Mason, “Making confident speaker verification decisions with minimal speech,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 6, pp. 1182–1192, Aug. 2010.

[21] M.-W. Mak, R. Hsiao, and B. Mak, “A comparison of various adaptation methods for speaker verification with limited enrollment data,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2006, vol. 1, pp. I–I.

[22] R. J. Vogt, C. J. Lustri, and S. Sridharan, “Factor analysis modelling for speaker verification with short utterances,” in *Proc. Odyssey*, 2008.

[23] A. Kanagasundaram, R. Vogt, D. B. Dean, S. Sridharan, and M. W. Mason, “i-vector based speaker recognition on short utterances,” in *Proc. 12th Annu. Conf. Int. Speech Commun. Assoc. (ISCA)*, 2011, pp. 2341–2344.

[24] M. Nosratighods, E. Ambikairajah, J. Epps, and M. J. Carey, “A segment selection technique for speaker verification,” *Speech Commun.*, vol. 52, no. 9, pp. 753–761, 2010.

[25] T. K. Moon, “The expectation-maximization algorithm,” *IEEE Signal Process. Mag.*, vol. 13, no. 6, pp. 47–60, Nov. 1996.

[26] J.-L. Gauvain and C.-H. Lee, “Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains,” *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

[27] M. K. Omar and J. W. Pelecanos, “Training universal background models for speaker recognition.” in *Proc. Odyssey*, 2010, pp. 52–57.

[28] J.-Y. Zhang, T. F. Zheng, J. Li, C.-H. Luo, and G.-L. Zhang, “Improved context-dependent acoustic modeling for continuous chinese speech recognition.” in *Proc. INTERSPEECH*, 2001, pp. 1617–1620.

[29] I. P. Association, *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge, U.K.: Cambridge Univ. Press, 1999.

[30] T. Schultz and A. Waibel, “Language-independent and language-adaptive acoustic modeling for speech recognition,” *Speech Commun.*, vol. 35, no. 1, pp. 31–51, 2001.

[31] À. Colomé, “Lexical activation in bilinguals’ speech production: Language-specific or language-independent?,” *J. Memory Lang.*, vol. 45, no. 4, pp. 721–736, 2001.

[32] H. Beigi, *Fundamentals of Speaker Recognition*. New York, NY, USA: Springer, 2011.

[33] C. Gong, *Research on Highly Distinguishable Speech Selection Methods in Speaker Recognition*. Beijing, China: Tsinghua University, 2014.

[34] N. Fatima, X.-J. Wu, T. F. Zheng, C.-H. Zhang, and G. Wang, “A universal phoneme-set based language independent short utterance speaker recognition,” in *Proc. 11th Nat. Conf. Man-Mach. Speech Commun. (NCMMSC’11)*, Xi’an, China, 2011, pp. 16–18.

[35] S. J. Young, J. J. Odell, and P. C. Woodland, “Tree-based state tying for high accuracy acoustic modelling,” in *Proc. Workshop Hum. Lang. Technol.*, 1994, pp. 307–312.

[36] A. W. Black and P. A. Taylor, “Automatically clustering similar units for unit selection in speech synthesis,” in *Proc. Eurospeech*, 1997, vol. 2, pp. 601–604.

[37] C. J. Leggetter and P. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models,” *Comput. Speech Lang.*, vol. 9, no. 2, pp. 171–185, 1995.

[38] M. J. Gales, “Maximum likelihood linear transformations for HMM-based speech recognition,” *Comput. Speech Lang.*, vol. 12, no. 2, pp. 75–98, 1998.

[39] A. Hall, “Methods for demonstrating resemblance in taxonomy and ecology,” *Nature*, vol. 214, pp. 830–831, 1967.

[40] B. Xiang and T. Berger, “Efficient text-independent speaker verification with structural Gaussian mixture models and neural network,” *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 447–456, Sep. 2003.

[41] J. A. Bilmes, “A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models,” *Int. Comput. Sci. Inst.*, vol. 4, no. 510, p. 126, 1998.

[42] A. Stolcke, S. S. Kajarekar, L. Ferrer, and E. Shrinberg, “Speaker recognition with session variability normalization based on MLLR adaptation transforms,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 7, pp. 1987–1998, Sep. 2007.

[43] 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. 2nd Int. Conf. Spoken Lang. Process. (ICSLP)*, 1992, vol. 92, pp. 599–602.

[44] C.-H. Zhang, L.-L. Wang, J. Jang, and T. F. Zheng, “A multimodel method for short-utterance speaker recognition,” in *Proc. Asia-Pac. Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, 2011.

[45] C. Zhang, X.-J. Wu, T. F. Zheng, L.-L. Wang, and C. Yin, “A K-phoneme-class based multi-model method for short utterance speaker recognition,” in *Proc. Asia-Pac. Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, 2012, vol. 20, no. 12, pp. 1–4.

[46] Z.-Y. Xiong, F. Zheng, W. Wu, and J. Li, “An automatic prompting texts selecting algorithm for DI-IFS balanced speech corpus,” in *Proc. Nat. Conf. Man-Mach. Speech Commun.*, 2003, pp. 252–256.

[47] S. Dobrisek, F. Mihelic, and N. Pavesic, “Acoustical modelling of phone transitions: Biphones and diphones-what are the differences?,” in *Proc. 6th Eur. Conf. Speech Commun. Technol.*, 1999, pp. 1307–1310.

[48] D. Wang, X.-Y. Zhu, and Y. Liu, "Multi-layer channel normalization for frequency-dynamic feature extraction," *J. Software*, vol. 12, no. 9, pp. p1523–1529, 2005.

[49] D. Povey *et al.*, "The kaldi speech recognition toolkit," in *Proc. IEEE Signal Process. Soc. Workshop Autom. Speech Recognit. Understand.*, Dec. 2011, Catalog No.: EPFL-CONF–192584.

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

[51] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. Eurospeech*, vol. 4, 1997, pp. 1895–1898.

Lantian Li received the B.Sc. degree from China University of Mining and Technology, Beijing, China, in 2013. He is currently pursuing the Ph.D. degree at the Center for Speech and Language Technology (CSLT), Tsinghua University, Beijing, China. His research interests include speaker recognition with machine learning methods.

Dong Wang (M'09) received the B.Sc. and M.Sc. degrees in computer science from Tsinghua University, Beijing, China, in 1999 and 2002, respectively, and the Ph.D. degree (supported by a Marie Curie fellowship) from CSTR, University of Edinburgh, Edinburgh, U.K., in 2010. He was with Oracle China from 2002 to 2004, and IBM China from 2004 to 2006. He joined CSTR, University of Edinburgh in 2006, as a Research Fellow. From 2010 to 2011, he was with EURECOM as a Postdoctoral Fellow, and from 2011 to 2012, was a Senior Research Scientist with Nuance. He is now an Assistant Professor with Tsinghua University.

Chenhao Zhang received the B.Sc. degree in computer science from Beijing University of Posts and Telecommunications, Beijing, China, in 2009. Since 2009, he has been with the Center for Speech and Language Technology, Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing, China. His research interests include speaker recognition, particularly with limited training/test data.

Thomas Fang Zheng (M'99–SM'06) received the Ph.D. degree in computer science and technology from Tsinghua University, Beijing, China, in 1997. He is now a Research Professor and the Director of the Center for Speech and Language Technologies, Tsinghua University. He has authored more than 250 papers. His research interests include speech and language processing.

Dr. Zheng plays active roles in a number of communities, including the Chinese Corpus Consortium (Council Chair), the Standing Committee of China¡s National Conference on Man-Machine Speech Communication (Chair), Subcommittee 2 on Human Biometrics Application of Technical Committee 100 on Security Protection Alarm Systems of Standardization Administration of China (Deputy Director), the Asia-Pacific Signal and Information Processing Association (APSIPA) (Vice-President and Distinguished Lecturer from 2012 to 2013), Chinese Information Processing Society of China (council member and Speech Information Subcommittee Chair), the Acoustical Society of China (council member), and the Phonetic Association of China (council member). He was an Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING and the *APSIPA Transactions on Signal and Information Processing*. He is on the Editorial Board of *Speech Communication, Journal of Signal and Information Processing*, *SpringerBriefs in Signal Processing*, and the *Journal of Chinese Information Processing*.