CSLT TECHNICAL REPORT-20150015 [Monday 24th August, 2015]

Improving Short Utterance Speaker Recognition by Modeling Speech Unit Classes

Chenhao Zhang[1], Dong Wang[1], Lantian Li[1] and Thomas Fang Zheng[1*]

*Correspondence:
fzheng@tsinghua.edu.cn
[1]Center for Speech and Language Technology, Research Institute of Information Technology, Tsinghua University, ROOM 1-303, BLDG FIT, 100084 Beijing, China
Full list of author information is available at the end of the article

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 as short as 2 seconds, the proposed subregion modeling offered a 23.64% 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.

Keywords: Short Utterance; Speaker Recognition; Subregion Model; Model Synthesis

1 Introduction

Speaker recognition (also named as speaker verification) aims to verify claimed identities of speakers. 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 speech signals are clear enough [1, 2, 3, 4, 5].

A popular approach to speaker recognition is the GMM-UBM framework [6, 7]. This approach involves a well-trained universal background model (UBM) to represent general 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 assumes that speaker and session variance distributes in two low-dimensional subspaces [9], the i-vector approach models speaker and session variance in a single low-dimensional subspace [10]. To improve the i-vector model, a multitude of normalization techniques

have been proposed, such as with-in class covariance normalization (WCCN) [11] and nuisance attribute projection (NAP) [2].

Recently deep learning has gained much success in multiple domains and caused extensive interests [12]. For speaker recognition, a very recent study applies DNN models trained for speech recognition to substitute UBMs, so that rich information in phones are employed to build more accurate models than GMMs that are trained in an unsupervised way [13, 14]. Additionally, DNNs have been utilized to extract speaker features [15, 16].

1.1 Challenge with short utterance

In spite of the great achievement, current speaker recognition systems perform well only if the enrollment and test data are abundant. In most 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, if not impossible, to collect sufficient data, for example in forensic applications. If the enrollment and test utterances are in the same text (so called 'text-dependent' task), short utterances would be not a big problem [17]; 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 a wealth of studies [18, 19, 20]. For instance, Vogt et al. reported that when the test speech was shortened from 20 seconds to 2 seconds, the performance in term of equal error rate (EER) increased sharply from 6.34% to 23.89% on a NIST SRE task [21]. Mak et al. showed that when the length of the test speech is less than 2 seconds, the EER was raised to as high as 35.00% [20]. Table 1 presents some results obtained in our study, where the enrollment data is sufficient and the test utterances vary from 300 to 2 seconds.

Table 1 Impact of the length of test utterances

Length (s)	300	20	10	5	2
EER (%)	6.34	8.87	12.15	16.99	23.89

1.2 Research on short utterance speaker recognition

The research on short utterance speaker recognition (SUSR) is still limited. In [19], the authors show that performance on short utterances can be improved by separating the speaker variation and the session variation in the framework of joint factor analysis (JFA). This work is extended in [22] 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 [23], 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 researches are based on test utterances that are of 5∼10 seconds. 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 seconds or 2 words in maximum, and dealing with the situation where both the test and enrollment utterance are short.

1.3 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 universal background model (UBM) and to enroll and test a particular speaker. Following the standard framework of Gaussian mixture model-universal back ground model (GMM-UBM), the characteristic of a particular speaker is modeled by a GMM. A commonly adopted GMM-UBM setup is to train an UBM by a pool of speech data involving a large number of speakers via the EM algorithm [24], and then a speaker's model is derived from the UBM the enrollment speech by MAP estimation [25] with only mean vectors being adapted. With this setup, the likelihood of a test utterance $x = \{x_t; t = 1, 2, ..., T\}$ 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 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 abundant, 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 enrollments/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.

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 not the prior-mismatch problem 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 not overlapped. 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, so impact the likelihood computation.

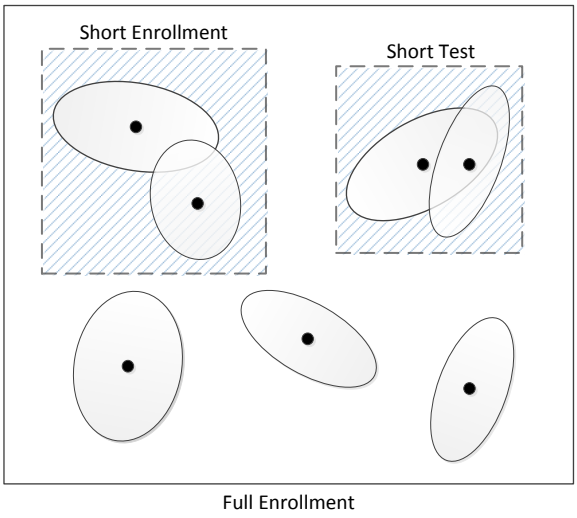

Figure 1 Mismatch between the model prior and the true distributions of enrollment/test speech signals.

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 the follows:

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

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:

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

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 role, but this is usually not satisfactory as the unsupervised learning does not leverage any external knowledge so the resulting model would be not 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 assign by the technique of forced phone alignment. This approach possesses several advantages. 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 scare 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 5; 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 2 discusses some related works, and 3 presents the subregion modeling, where we assume that the enrollment data is sufficient. Section 4 presents the model synthesis approach to deal with speech units with limited enrollment data. Section 5 describes the experiments, and the entire paper is concluded in Section 6.

2 Related work

The idea of employing phone information in speaker recognition has been investigated by other researchers, particularly with the DNN-based method proposed by Lei and colleagues [14]. The difference is that they use the DNN-based phone posteriors to replace GMM-based class posteriors to train i-vector models, while we use phone posteriors or alignments to partition acoustic space into subregions, and each subregion is still modeled by a GMM. Another difference is that Lei's method [14] employs the phone knowledge only in model training, while our method employs it in both training and test.

Using phone knowledge is also a unique advantage when comparing the subregion model to the i-vector model in SUSR. It is well known that the i-vector model posses some advantage when dealing with short enrollment/test utterances [22], due to its nature of sharing statistical strength among different acoustic regions. However this model is purely unsupervised and does not utilize any phone knowledge. This

problem can be mitigated to some extent by the DNN-based method discussed above [14], however the phone knowledge conveyed by DNNs only exists in UBM training. The subregion model proposed in this work, in contrast, utilizes the phone knowledge in both model training, speaker enrollment and test. We believe there are some methods that can be used to combine these two different approaches but leave the investigation as future work.

3 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.

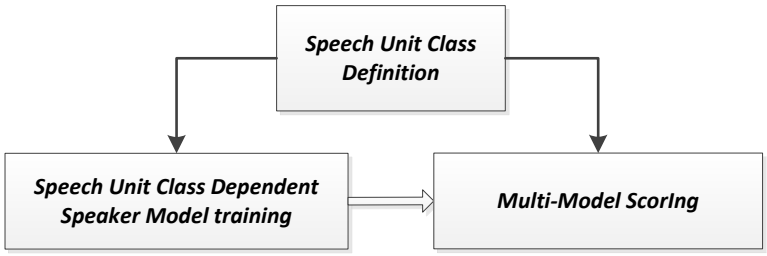

Figure 2 The Framework of the subregion modeling

3.1 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 [26]. Although language-independent speech units can be defined, e.g., through the International Phonetic Association (IPA) [27] and multi-lingual speaker/speech recognition systems [28, 29], 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 [26]. 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 2, where {_a, _o, _e, _i, _u, _v} are zero initials and appear in non-initial syllables [26].

Among the IFs, Finals have been found conveying more spectral information than Initials [30]. Better speaker recognition performance therefore can be obtained by selecting speech segments corresponding to Finals only. To verify this conjecture, we built three GMM-UBM speaker recognition systems, with speech segments of Initials, Finals and all the IFs, respectively. The experiments were conducted on SUD12, a Chinese SUSR database recorded at the Tsinghua University (details of

Table 2 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

the database are given in Section 5). An off-the-shelf speech recognition system trained on a large database and with the same IF set was used to segment the speech signals into IF segments. The results in EER are shown in Table 3. It is clear to see that the system based on the Finals delivers much better performance than that based on the Initials and the entire IF set. Based on this result, we choose IFs as the speech units in this work, but only the speech segments of Finals are used to build systems. In other words, the Finals are the effective speech units when constructing subregion models in this study.

Table 3 EERs with different IF sets

Data Type	EER (%)
All IFs	7.16
Initials	40.25
Finals	5.86

3.2 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.

3.2.1 ***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 [31], which is based on tongue's height and backness information of the speech units in the IPA definition.

3.2.2 ***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 [32] and unit selection in speech synthesis [33], the greedy merge of similar classes used in maximum likelihood linear

regression (MLLR) [34, 35]. Most of these approaches try various possible merge schemes and select the best one that leads to the highest probability on training data. In this study, we develop a vector quantization (VQ) method based on the K-means algorithm [36] 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. Because the clustering method itself is not the main focus of this work, we believe this simple algorithm is sufficient for our purpose. 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 channels and genders.
- Let N denote the number of Finals. Collecting 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 $\lambda_i = \{\pi_k, \mu_{i,k}, \Sigma_k : k = 1, ..., K\}$. Note that only $\{\mu_{i,k} : k = 1, ..., K\}$ are Final-dependent.
- Define the distance of two Final-dependent UBMs based on the symmetric Kullback-Leibler (KL) divergence [37], given by:

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

  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 [36] is conducted to cluster the N Final-dependent UBMs into C clusters, with the distance measure set to the KL divergence.

3.3 Subregion modeling based on speech unit classes

Denote the speech unit classes (Final clusters) by $\{\text{SUC-}c := 1, ..., C\}$. 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}. Train a global UBM with the entire training dataset, by employing the expectation-maximization (EM) algorithm [24, 38].
- 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 {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 [25] and with the speech data assigned to SUC-c.
- 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.

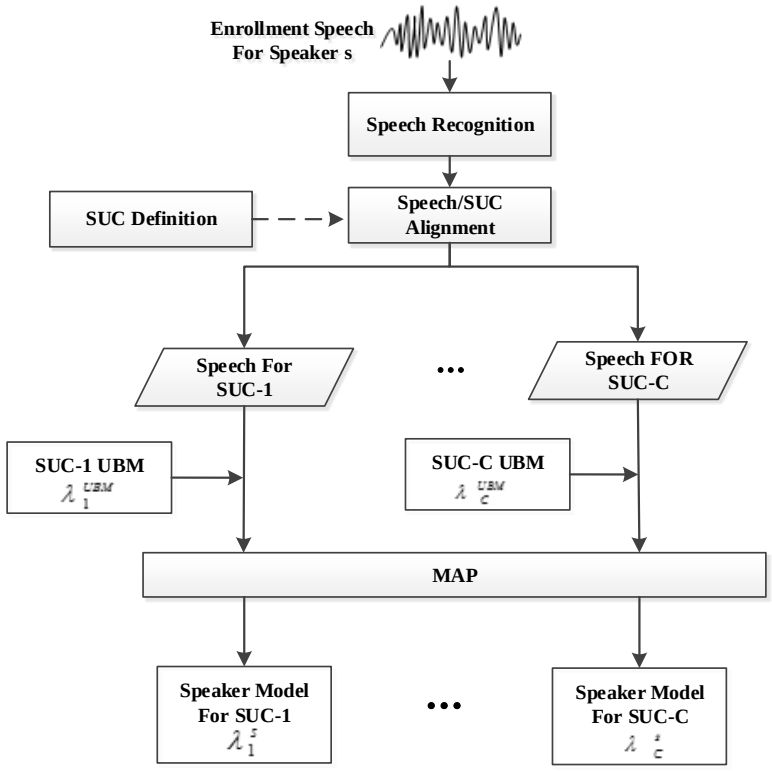

Figure 3 Speaker-dependent subregion model training. 'SUC' stands for speech unit class.

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 and 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.

3.4 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 denote 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 $\lambda^s_{c(l)}$ and the subregion UBM $\lambda^{UBM}_{c(l)}$, where s denotes the speaker. This is formulated by:

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

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

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

4 Speaker model synthesis

The subregion modeling presented in the previous section assumes distributes, 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 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 adapt speaker-dependent models (subregion GMMs) from speaker-independent models (subregion UBMs), and the relationship can be represented by a linear transform. These transforms then can 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.

4.1 Maximum likelihood linear regression

The maximum likelihood linear regression (MLLR) [34, 39] was first proposed by the Cambridge group to deal with channel mismatch and speaker variability in speech

recognition. Given a GMM $\lambda = (\pi_k, \mu_k, \Sigma_k : k = 1, 2, ..., K)$ and a speech segment X, the MLLR seeks for 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

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

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, ${G_i}^{-1}$ is 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 $x(t)$ 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.

4.2 Model synthesis based on subregion UBMs

With the MLLR technique, a transforms $L_{i,j}$ 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 classes 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:

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

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.

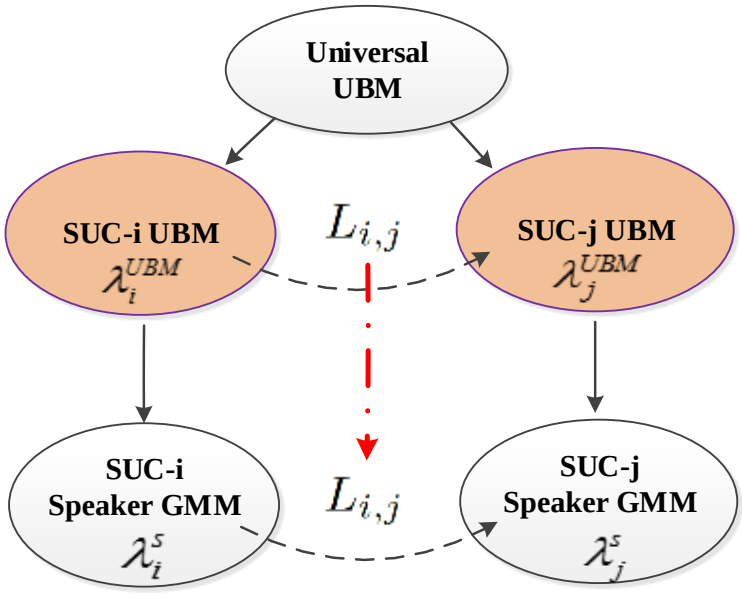

Figure 4 Illustration of model synthesis based on subregion UBMs. 'SUC' stands for speech unit class.

4.3 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 assumption is over strong, as different speakers may exhibit clear different characteristics when moving from one pronunciation to another. We propose speaker-dependent transforms based on cohort sets.

A cohort set [40] 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 the 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 $\{\lambda_c^h : c = 1, 2, ..., C\}$, where h

indexes the speaker and c indexes the subregion. Secondly the MLLR transform was estimated for each speaker h between each subregion pairs (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:

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

where k indexes the Gaussian components.

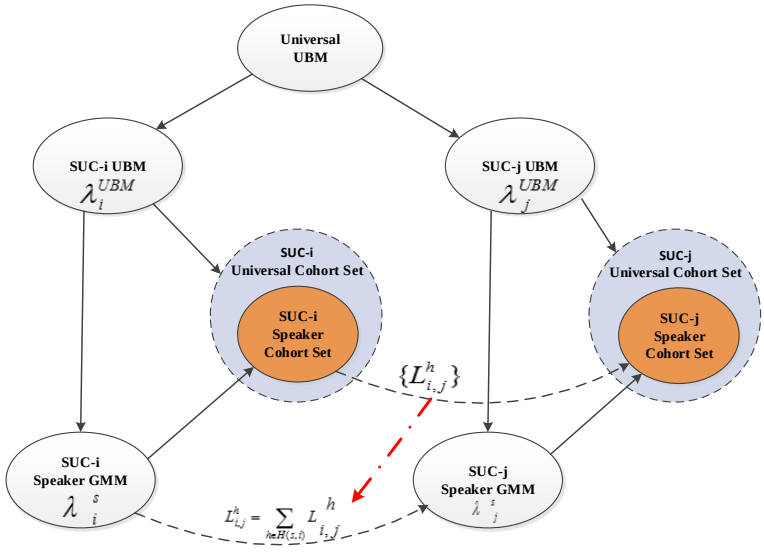

Figure 5 The illustration of model synthesis based on cohort speakers. 'SUC' stands for speech unit class.

5 Experiment

5.1 Database

5.1.1 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, may introduce artifacts when cutting continuous speech signals. We therefore decided to design and recorded a database that is suitable for

SUSR research and publish it for research usage[1]. The database was named as "SUD12" [41, 42], 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 neural. There are in total 30 male speakers and 30 female speakers, and all the utterances are in Standard Chinese. The sampling rate is 16 kHz, and the sampling precision is 16 bits.

The enrollment dataset involves 100 Chinese sentences, each of which contains 15 $\sim$ 30 Chinese characters, and the average length of effective speech signals is about 10 seconds. These sentences were selected by the ELFU algorithm [43] from $5,000$ sentences in the news domain taken from the People's Daily, with the objective to maximize the di-IF coverage [44]. 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 4.

Table 4 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

The test dataset of SUD12 involves 63 short utterances, which covered all the Finals in Standard Chinese. The lengths of the recordings are not more than 2 seconds. The distribution is shown in Table 5.

Table 5 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

5.1.2 Database for UBM training (863DB)

The speech data used to train the UBMs and subregion UBMs were chosen from the 863 Chinese speech corpus [45]. 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 80 males and 80 females from

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

the 863 corpus, and for each speaker, there are 75 speech utterances, and the length of the speech signals is 17 hours in total. This dataset is denoted by 863DB for convenience.

5.1.3 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, 100 Standard Chinese sentences were recorded, and each utterance involves 10 seconds of effective speech signals approximately. This database is denoted by dEarDB.

5.2 Experimental conditions

The acoustic feature is the conventional 32-dimensional Mel frequency cepstral coefficients (MFCC), which involves 16-dimensional static components plus the first order derivatives. Note that a simple energy-based voice activity detection (VAD) has been performed before the feature extraction, and the cepstral mean normalization (CMN) [46] is applied as a post-process to reduce the impact of channel mismatch.

We choose the conventional GMM-UBM approach to construct the baseline system. The UBM consists of $1,024$ Gaussian components and is 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 is employed to conduct the evaluation. With the enrollment data, the speaker GMMs are derived from the UBM by MAP. The test result on the SUD12 test set is 29.78% in EER. This is a reasonable performance for SUSR that involves short utterances less than 2 seconds [19, 20].

5.3 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 [31]. 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 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 both too small and too large values lead to suboptimal performance, and the optimal setting in our experiment is C=6. Table 6 shows the derived unit classes with this configure. It can be seen that the clustering result is reasonable at least intuitively.

The results in terms of EER are presented in Table 7, where 'GMM-UBM' is the baseline system, and 'SBM-KW' and 'SBM-DD' are subregion systems with

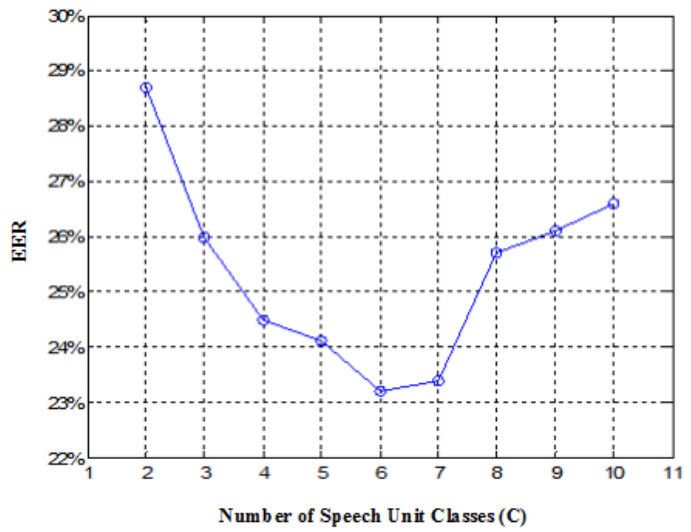

Figure 6 EERs with different numbers of speech unit classes in data-driven clustering.

Table 6 Speech unit classes derived in data-driven way

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

the knowledge-based and data-driven speech unit clustering, respectively. 'EERR' stands for relative EER reduction. 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 alarm (incorrect acceptance) probability and the vertical axis represents the miss probability (incorrect rejection) [47]. It can be seen that the systems based on subregion modeling outperforms 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 characteristics of real data, and the balance of data over the resultant speech unit classes may have lead to more robust subregion models.

One may argue that the comparison in Table 7 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.

Table 7 Performance of subregion modeling

System	EER (%)	EERR (%)
GMM-UBM (baseline)	29.78	–
SBM-KW	25.80	13.36
SBM-DD	22.74	23.64

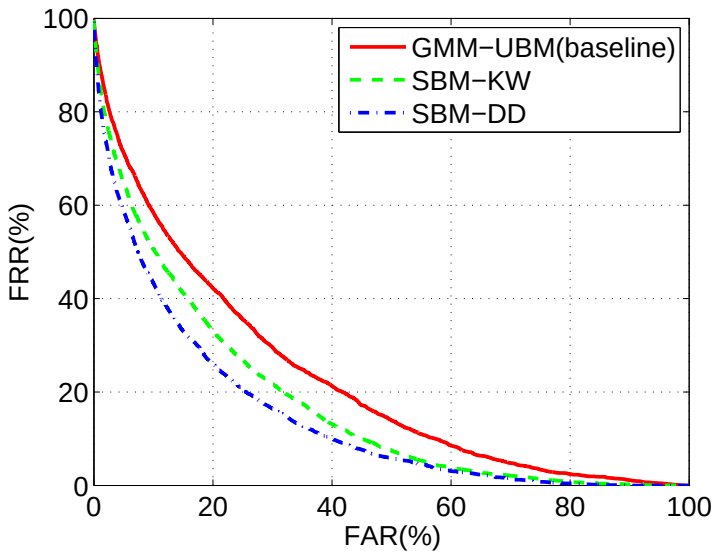

Figure 7 The DET curves with the GMM-UBM baseline and subregion models.

5.4 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 of Table 6, and simulate data-sparse speech unit classes by discarding the speech segments assigned to the class.

5.4.1 Synthesis based on subregion UBMs

We study the model synthesis approach based on subregion UBMs. The results are shown in Table 8, 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 'NULL' presents the results without any model synthesis. It can be seen that with the model synthesis, the performance is generally improved compared with the baseline system.

Table 8 Results with model synthesis based on subregion UBMs

EER (%)	SBB1	SBB2	SBB3	SBB4	SBB5	SBB6	NULL
SBS1	–	26.61	27.09	27.08	27.29	26.68	28.94
SBS2	27.12	–	27.06	27.04	27.27	26.63	30.26
SBS3	27.21	27.18	–	27.47	26.62	26.98	29.92
SBS4	27.02	27.56	27.11	–	27.31	26.17	29.55
SBS5	27.10	27.15	26.79	27.11	–	26.69	29.43
SBS6	26.94	27.52	27.01	27.25	27.32	–	29.12

5.4.2 Synthesis based on cohort speakers

As discussed in Section 4, 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 4-th speech unit class as the data-sparse class and synthesize the subregion model from the model of the 3-th speech unit class.

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 20. This optimal value is used in the rest of the experiments.

Table 9 presents the results with the MLLR-based model synthesis, where the row 'NO-MLLR' present the system without any treatment for data-sparse speech unit classes. 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. The values in the 'EERR' column are EER reductions compared with the 'NO-MLLR' system. 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 outperforms the subregion UBM-based synthesis.

Table 9 Results with model synthesis

System	EER (%)	EERR (%)
SBM-DD	22.74	-
NO-MLLR (baseline)	29.55	-
MLLR-UBM	27.11	8.26
MLLR-COHORT	24.33	17.66

6 Conclusions

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 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.

Author details

[1]Center for Speech and Language Technology, Research Institute of Information Technology, Tsinghua University, ROOM 1-303, BLDG FIT, 100084 Beijing, China. [2]Department of Computer Science and Technology, Tsinghua University, ROOM 1-303, BLDG FIT, 100084 Beijing, China.

References

1. J. P. Campbell Jr, "Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 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," *Computer Speech & Language*, vol. 20, no. 2, pp. 210–229, 2006.
3. F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, vol. 2004, pp. 430–451, 2004.
4. T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech communication*, vol. 52, no. 1, pp. 12–40, 2010.
5. C. S. Greenberg, V. M. Stanford, A. F. Martin, M. Yadagiri, G. R. Doddington, J. J. Godfrey, and J. Hernandez-Cordero, "The 2012 nist speaker recognition evaluation," in *Proc. INTERSPEECH'13*, 2013, pp. 1971–1975.
6. D. A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech communication*, 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," *Digital Signal Processing*, vol. 10, no. 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 Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.
9. P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.
10. N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.
11. A. O. Hatch, S. S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for svm-based speaker recognition," in *Proc. INTERSPEECH'06*, 2006.
12. L. Deng and D. Yu, *DEEP LEARNING: Methods and Applications*. Foundations and Trends in Signal Processing, January 2014.
13. P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet, and J. Alam, "Deep neural networks for extracting baum-welch statistics for speaker recognition," 2014.
14. Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014*. IEEE, 2014, pp. 1695–1699.
15. V. Ehsan, L. Xin, M. Erik, L. M. Ignacio, and G.-D. Javier, "Deep neural networks for small footprint text-dependent speaker verification," in *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP),2014*. IEEE, 2014, pp. 4052–4056.
16. 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.
17. A. Larcher, K.-A. Lee, B. Ma, and H. Li, "Rsr2015: Database for text-dependent speaker verification using multiple pass-phrases," in *Proc. INTERSPEECH'12*, 2012.
18. M. McLaren, D. Matrouf, R. Vogt, and J.-F. Bonastre, "Applying svms and weight-based factor analysis to unsupervised adaptation for speaker verification," *Computer Speech & Language*, vol. 25, no. 2, pp. 327–340, 2011.
19. R. J. Vogt, C. J. Lustri, and S. Sridharan, "Factor analysis modelling for speaker verification with short utterances," in *The Speaker and Language Recognition Workshop*. IEEE, 2008.
20. M.-W. Mak, R. Hsiao, and B. Mak, "A comparison of various adaptation methods for speaker verification with limited enrollment data," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2008*, vol. 1. IEEE, 2006, pp. I–I.
21. R. Vogt, S. Sridharan, and M. Mason, "Making confident speaker verification decisions with minimal speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1182–1192, 2010.

22. A. Kanagasundaram, R. Vogt, D. B. Dean, S. Sridharan, and M. W. Mason, "i-vector based speaker recognition on short utterances," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association*. International Speech Communication Association (ISCA), 2011, pp. 2341–2344.
23. M. Nosratighods, E. Ambikairajah, J. Epps, and M. J. Carey, "A segment selection technique for speaker verification," *Speech Communication*, vol. 52, no. 9, pp. 753–761, 2010.
24. T. K. Moon, "The expectation-maximization algorithm," *Signal processing magazine, IEEE*, vol. 13, no. 6, pp. 47–60, 1996.
25. J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Transactions on Speech and audio processing*, vol. 2, no. 2, pp. 291–298, 1994.
26. 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'01*, 2001, pp. 1617–1620.
27. I. P. Association, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.
28. T. Schultz and A. Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, no. 1, pp. 31–51, 2001.
29. À. Colomé, "Lexical activation in bilinguals' speech production: Language-specific or language-independent?" *Journal of memory and language*, vol. 45, no. 4, pp. 721–736, 2001.
30. H. Beigi, *Fundamentals of speaker recognition*. Springer, 2011.
31. 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 *11th National Conference on Man-Machine Speech Communication (NCMMSC'11), Xi'an, China*, 2011, pp. 16–18.
32. S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proceedings of the workshop on Human Language Technology*. Association for Computational Linguistics, 1994, pp. 307–312.
33. A. W. Black and P. A. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," 1997.
34. C. J. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech & Language*, vol. 9, no. 2, pp. 171–185, 1995.
35. M. J. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer speech & language*, vol. 12, no. 2, pp. 75–98, 1998.
36. A. Hall, "Methods for demonstrating resemblance in taxonomy and ecology," *Nature*, vol. 214, pp. 830–831, 1967.
37. B. Xiang and T. Berger, "Efficient text-independent speaker verification with structural gaussian mixture models and neural network," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 447–456, 2003.
38. J. A. Bilmes, "A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models," *International Computer Science Institute*, vol. 4, no. 510, p. 126, 1998.
39. A. Stolcke, S. S. Kajarekar, L. Ferrer, and E. Shrinberg, "Speaker recognition with session variability normalization based on mllr adaptation transforms," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 1987–1998, 2007.
40. 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 *ICSLP*, vol. 92, 1992, pp. 599–602.
41. C.-H. Zhang, L.-L. Wang, J. Jang, and T. F. Zheng, "A multimodel method for short-utterance speaker recognition," in *Asia-Pacific Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2011.
42. 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 *Asia-Pacific Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC)*, vol. 20, no. 12, 2012, pp. 1–4.
43. Z.-Y. Xiong, F. Zheng, W. Wu, and J. Li, "An automatic prompting texts selecting algorithm for di-ifs balanced speech corpus," in *National Conference on Man-Machine Speech Communications*, 2003, pp. 252–256.
44. S. Dobrisek, F. Mihelic, and N. Pavesic, "Acoustical modelling of phone transitions: biphones and diphones-what are the differences?" in *Sixth European Conference on Speech Communication and Technology*, 1999.
45. D. Wang, X.-Y. Zhu, and Y. Liu, "Multi-layer channel normalization for frequency-dynamic feature extraction," *Journal of Software*, vol. 12, no. 9, pp. p1523–1529, 2005.
46. S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, no. 2, pp. 254–272, 1981.
47. A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The det curve in assessment of detection task performance," DTIC Document, Tech. Rep., 1997.

Acknowledgement

This work was supported by the National Natural Science Foundation of China under Grant No.61271389/61371136 and the National Basic Research Program(973 Program) of China under Grant No.2013CB329302.