A Multi-Model Method for Short-Utterance Speaker Recognition

Chenhao Zhang[1], Xiaojun Wu[1], Linlin Wang[1], Gang Wang[1], Jyh-Shing Roger Jang[2], and Thomas Fang Zheng[1*]
[1] Center for Speech and Language Technologies, Division of Technical Innovation and Development,
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology, Tsinghua University, Beijing, 100084
[2] Department of Computer Science, Tsing Hua University, Hsin-chu
E-mail: {zhangchh, wangll, wanggang} @cslt.riit.tsinghua.edu.cn, xjwu@tsinghua.edu.cn, jang@cs.nthu.edu.tw
* Corresponding Author: fzheng@tsinghua.edu.cn Tel/Fax: +86-10-62796393

***Abstract*—The length of the test speech greatly influences the performance of GMM-UBM based text-independent speaker recognition system, for example when the length of valid speech is as short as 1~5 seconds, the performance decreases significantly because the GMM-UBM based speaker recognition method is a statistical one, of which sufficient data is the foundation. Considering that the use of text information will be helpful to speaker recognition, a multi-model method is proposed to improve short-utterance speaker recognition (SUSR) in Chinese. We build a few phoneme class models for each speaker to represent different parts of the characteristic space and fuse the scores to fit the test data on the models with the purpose of increasing the matching degree between training models and test utterance. Experimental results showed that the proposed method achieved a relative EER reduction of about 26% compared with the traditional GMM-UBM method.**

I. Introduction

Short-utterance speaker recognition (SUSR) is becoming more and more important nowadays in tasks where the lengths of valid speech in test utterances are very short. It can be applied in many fields of real environments, such as security check, banking business and so on. Using voice to recognize the identities can be more comfortable and convenient for the users and more cost-effective compared with other biometrics-based identification techniques. In some situations, the speaker's identity should be verified only by one or two words. In other situations, speaker recognition with short utterances can provide a better user experience. So SUSR is worth being researched.

Over the past few years, many approaches have been applied in speaker recognition, and people have done plenty of researches. Most of the approaches are based on GMM-UBM [2] or GMM-SVM [3]. Mel frequency cepstrum coefficient (MFCC) [4] is one of the most commonly used features. Feature vectors extracted from a test utterance (speaker-unknown) are used to calculate the likelihood scores [5] against all target speaker models and these scores are used to judge which speaker fits the test utterance best. Most research work is based on the precondition of enough test speech data. It is shown that the length of the test data is a great factor that influences the performance. [6] shows some results based on the NIST SRE 2005 [5] database (See Table I), and the evaluation parameters are Equal Error Rate (EER) and minimum Detection Cost Function (minDCF) [5]. Furthermore, if the length is less than 2 seconds, EER rises to about 35%.

TABLE I
THE EFFECT OF SHORTENED LENGTH TEST UTTERANCES ON SPEAKER RECOGNITION PERFORMANCE[10]

System	EER (%)	minDCF (10^{-2})
Baseline	6.34	2.93
20 seconds	8.87	3.91
10 seconds	12.15	4.89
5 seconds	16.99	6.16
2 seconds	23.89	7.94

Due to the restriction of practical applications, it is difficult to manage users to record long speech. Moreover, traditional algorithms do not perform smoothly and efficiently in this short utterance situation.

In order to improve the performance of SUSR, some approaches have been proposed. Factor analysis subspace estimation [7] based on the JFA algorithm is used for developing a joint factor analysis in GMM speaker verification system, and it is shown that the performance of this system is improved in short utterance situation. [8, 9] try to improve the performance by selecting segments with higher discriminability on speaker characteristics to do speaker recognition. The weighted bilateral scoring method is used to enhance the performance of speaker recognition in the scoring domain [10]. However, most of the approaches above focus on the lengths between 5~10s. There are still challenges especially when the speech is shorter. In general, the GMM-UBM method builds a Gaussian mixture model for each target speaker, which represents the entire probability distribution of all training data. When the test utterance is very short, the test data only represents partial aspect of the speaker characteristics. However, the other parts of the space weaken the discriminability of the right part since the speaker model is trained entirely from all data. The shorter the length is, the stronger this effect will be. In a word, the discriminability is decreased because entire speaker model is adapted, and the performance of text-independent GMM-UBM system degrades intensely. A more accurate model would strengthen the discriminability and fit the test utterance better. In this paper, we focus on the SUSR in Chinese on the condition that the lengths of test utterances are less than 2 seconds and the training data is enough. A method for text-independent SUSR is proposed to build multi-phoneme class models of target speakers associated with a decision-making process. Its basic idea is to build a class of models for each target speaker. Each model represents a part of the entire speaker characteristics. Therefore the matches between the training and the testing situations are strengthened by calculating the distances between the test utterances and the models. Because the short utterances only contain one or twosyllables, phoneme information should be sufficiently utilized for the purpose of increasing the discriminability. The speaker models are built with the phoneme information, and the test utterance is scored against the corresponding phoneme model to increase the discriminability. The system is constructed as follows: First, a Chinese phoneme recognizer is built with speech recognition technology to obtain the phoneme information, then define the phoneme classes for all the phonemes and train phoneme class UBM (PCUBM) models. Second, multiple GMM models are derived from PCUBMs for each target speaker. Third, the test utterance is used to calculate the likelihood score against the multiple models. In other words, this method converts the current text-independent speaker recognition into text-dependent speaker recognition in some sense. A reduction is expected in EER compared to those methods based on GMM-UBM.

This paper is organized as follows. In Section II, the phoneme classes are defined and the SUSR framework based on multi-models is described in detail. The experimental settings, results and analysis are given in Section III. Conclusions and future work are presented in Section IV.

II. SUSR FRAMEWORK BASED ON MULTI-PHONEME-MODELS

A. Phoneme Classes Definition

In this paper, the utterances are in Chinese. The phoneme classes are defined according to the pronunciation characteristics. The reasons of choosing phoneme classes rather than phonemes themselves are that: first, some phonemes in Chinese pronunciation carry similar speaker characteristic. They share the similar Gaussian mixtures in the speaker property space and can be clustered into the same classes. Second, training models for each phoneme is not appropriate because it easily leads to the data sparsity problem since there are so many phonemes in Chinese pronunciation. Certain studies showed that vowels were proved to carry across them significant information about the speaker's identity, so in this paper the phoneme classes are defined based on the vowels in Chinese.

Let $\{P_1, P_2, \cdots, P_N\}$ denote the entire phoneme set in Chinese pronunciation, where N is the total number of Chinese phonemes and P_i denote one special phoneme. Phoneme classes are defined as:

$$PC_j : P_{j,1}, \cdots, P_{j,K_j}, j \in [1, n], \sum_{j=1}^{n} K_j = N$$

where PC_j is the j-th phoneme class, and there are n phoneme classes. For PC_j, it contains K_j phonemes whose acoustic characteristic are similar, and $P_{j,k}$ denotes the k-th phoneme in PC_j.

B. The Phoneme Recognizer for Phoneme Classes

The phoneme recognizer is built using Hidden Markov Model (HMM)[11] with MFCC features to recognize XIFs in Chinese pronunciation. The recognizer parameters are illustrated in Section III.

The training speeches are separated into phoneme segments according to the time labels given by the phoneme recognizer, and the phoneme segments are then grouped into several phoneme classes of segments. Since acoustically similar phonemes belong to the same class, recognition errors among these phonemes won't lead to errors in segment grouping.

C. SUSR Framework

The proposed SUSR framework based on multi-model is established for exploring which parts matchthe short test utterance best as shown below, and it contains a training phase and a test phase:

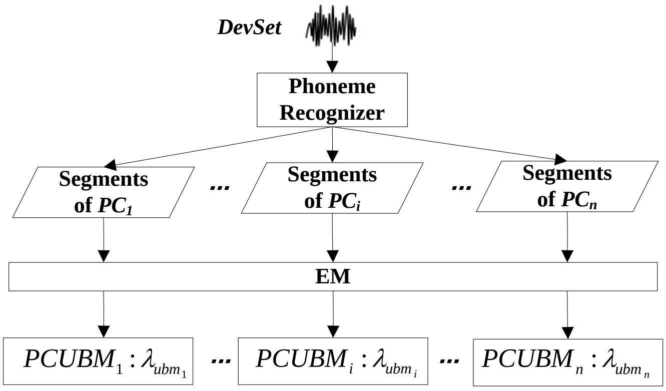

Fig. 1 Phoneme Class UBM Training Phase

The purpose of *PCUBM* training is to construct the basic standards for each phoneme class model. First, the phoneme segments in the utterances in the development set are obtained and grouped with the help of the phoneme recognizer, so the data are divided into n classes. Then the *PCUBM* model for PC_i is trained by expectation-maximization (EM) [12] algorithm with the segments of the i-th phoneme class, and λ_{ubm_i} denotes the model parameters of $PCUBM_i$, where i ranges from 1 to n.

Assume there are M target speakers. For the target speaker m, n phoneme class dependent models are constructed to replace the conventional speaker GMM model. Segments of $PC_{m,i}$ in Fig. 2 mean the i-th phoneme class data from Speaker m, and the i-th speaker model is self-adapted from $PCUBM_i$ by using the Maximum *a posteriori* (MAP) [13] algorithm. $\lambda_{m,i}$ denotes the model parameters of the i-th phoneme model for Speaker m, where m is from 1 to M.

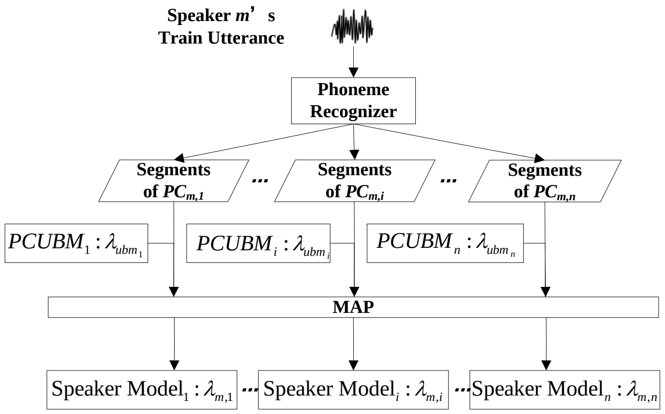
Speaker m's Train Utterance
Phoneme Recognizer
Segments of $PC_{m,1}$
...
Segments of $PC_{m,i}$
...
Segments of $PC_{m,n}$
$PCUBM_1 : \lambda_{ubm_1}$
$PCUBM_i : \lambda_{ubm_i}$
$PCUBM_n : \lambda_{ubm_n}$
MAP
Speaker Model$_1 : \lambda_{m,1}$
...
Speaker Model$_i : \lambda_{m,i}$
...
Speaker Model$_n : \lambda_{m,n}$

Fig. 2 Speaker Phoneme Class Model Training for Speaker *m*

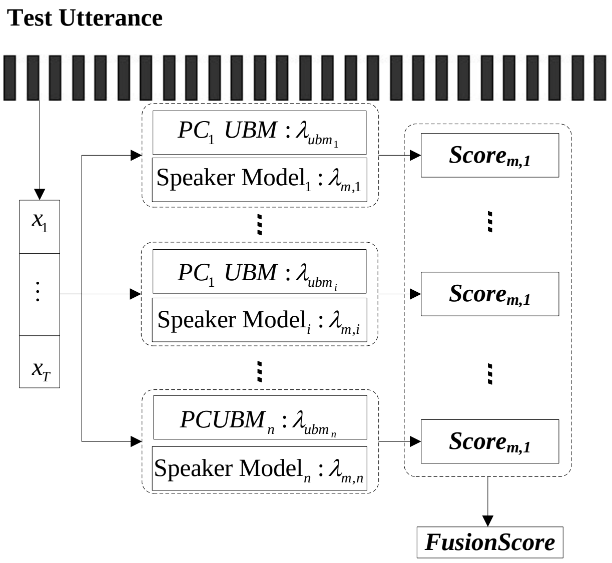
Test Utterance
x_1
⋮
x_T
PC_1 $UBM : \lambda_{ubm_1}$
Speaker Model$_1 : \lambda_{m,1}$
PC_1 $UBM : \lambda_{ubm_i}$
Speaker Model$_i : \lambda_{m,i}$
$PCUBM_n : \lambda_{ubm_n}$
Speaker Model$_n : \lambda_{m,n}$
$Score_{m,1}$
$Score_{m,1}$
$Score_{m,1}$
FusionScore

Fig. 3 Testing on Speaker *m*

Based on the above introduction for the training phase, the testing phase is straightforward. When the MFCC features of the test utterance are extracted, x_i denotes the MFCC feature of frame *i*, where *i* is from 1 to *T*. The decision score is fused from the scores which are calculated on the speaker *PC* models and PCUBM models as follows:

$$For\ \text{speaker}\ m: \quad i \in [1, n], m \in [1, M]$$

$$Score_{m,i} = \frac{1}{T}\sum_{j=1}^{T}\Big[\log p(x_j \mid \lambda_{m,i}) - \log p(x_j \mid \lambda_{ubm_i})\Big]$$

$$FusionScore_m = Fusion(Score_{m,1}, \ldots, Score_{m,n})$$

For a test utterance, the *n* scores will be obtained against the speaker *m*, and the fusion of scores can be either a voting method or any other one.

So the SUSR framework is illustrated as above, and there are some assumptions that need to be illustrated. First, all the phoneme class models should be well trained on account of the short test utterance situation, that is to say the training data should be enough, and it is an absolutely necessary and appropriate assumption. Second, as the focus point is to solve the problem of short utterances, so the speech data should be avoided from with unusual background noise, speaking styles and channel information.

III. Experimental Results and Analysis

A. Experimental data and set up

The SUSR experiments were performed on the short test utterance (STU) database which was specially constructed for SUSR. The short speech which is taken from a long sentence is not a good choice considering the co-articulation problem and the cut point inaccuracy problem.

The STU database consists of 163 utterances by 57 Chinese speakers (29 males and 28 females) and was recorded in clean environments by microphone in Mandarin Chinese. The 163 utterances are separated into a training part and a test part. The training part consists of 100 long sentences covering all vowels in Chinese pronunciation, and the texts of sentences were designed to be phoneme-balanced in Chinese pronunciation for the purpose of training all phoneme class models [13]. The test part consists of 63 short utterances, and the length of each test utterance is less than 2 seconds and the distribution of lengths is listed in Table II. All the utterances are in 16 kHz sampling rate with 16-bit width.

Table II
The distribution of the length of short test utterances

Length (s)	Number	Percent (%)
< 0.5	38	60.3
0.5 ~ 1.0	15	23.8
1.0 ~ 2.0	10	15.9

As illustrated in Section II, a Chinese phoneme recognizer was used to obtain the phoneme sequence. 50 hours SONY Chinese mandarin speech was used to train it. The feature that the recognizer used is MFCC feature, which contains 12-dimensional MFCC coefficients, and their acceleration coefficients, delta coefficients, energy and zero static coefficients. There are 65 3-state left-right no skip sub-models, and as explained in Section II, a rough model is acceptable, so each state consists of only 16 Gaussian mixtures.

The 38 Chinese vowels are separated into 6 phoneme classes, based on the Chinese vowel /a/, /e/, /i/, /o/, /u/, /v/ according to the expert knowledge.

The baseline system was a speaker recognition system based on the conventional GMM-UBM. The UBM consisted of 1,024 mixture components and was trained from the 863 CSL Corpus [14]. For the MAP training [15], only mean vectors were adapted with a relevance factor of 16. The cepstral mean subtraction in feature-domain was applied. These parameters were also used in the SUSR system. A voting method was used in the fusion of the scores of each phoneme class.

Feature extraction in the speaker recognition phase was performed on a 20 millisecond frame every 10 milliseconds. The pre-emphasis coefficient was 0.97 and Hamming windowing was utilized in each frame. 16-dimensional MFCC features were extracted from the training and test utterances with 30 triangular Mel filters in the MFCC calculation. For each frame, we extracted 32-dimentional feature vectors which were formed of the MFCC coefficients and their first derivative.

B. Experimental Results and Analysis

EER and minDCF were used to evaluate the performance of speaker recognition, and the parameters of minDCF were the same as in [5]. Results are listed in Table III, where *k* is the number of phoneme models that are scored against.

TABLE III
SPEAKER RECOGNITION PERFORMANCE
WITH DIFFERENT VALUE OF K IN MULTI-MODEL METHOD

k	EER (%)	minDCF (10^{-2})
1	28.99	9.98
2	28.35	9.91
3	28.12	9.89
4	27.90	9.87
5	27.74	9.83
6	27.73	9.83

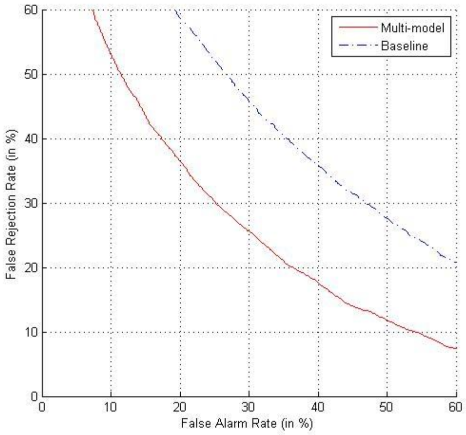

Fig. 4 DET curves of multi-model method vs GMM-UBM

The EER of the baseline system was 37.8%. Compared with the traditional GMM-UBM method, the multi-model method achieved a relative EER reduction of 26.64% (from 37.8% to 27.73%), and it can be seen that the EER basically did not change when *k* was larger than 5, which is evidence that the proposed method reached its upper bound. The DET curves of the systems are illustrated in Fig. 4. The results above verified the better capability of speaker recogniton with multi-modeling over GMM-UBM, and the voting fusion method provided a prominent reduction on EER. Because the test utterance is very short, it may only contain one or two syllables, that is to say, the test utterance should match one or two of the phoneme models best. The recognition result is obtained from the fusion of the scores that calculated against the speaker's multi-models, which can decrease the influence of error from the phoneme recognition. The experimental results showed that our method emphasized the match between the train models and the test utterances.

IV. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a text-independent short utterance speaker recognition method based on phoneme recognition and multi-models. The experimental results showed that our approach achieved a better result than the traditional GMM-UBM method in the short utterances situation.

There can be some further improvements on our method. The phoneme classes can be defined more accurately by though data-driven rather than expert knowledge. Other fusion method can be applied especially for giving weights to the phoneme classes. In addition, our method assumes that all the phoneme class models are trained enough, so using model self-adaption and mapping phoneme class models to apply the method can be further investigated when some phoneme class are not well trained.

REFERENCES

[1] J. P. Campbell. Speaker recognition: A tutorial. Proceedings of the IEEE, September 1997, vol. 85, pp. 1437--1462

[2] D. A. Reynolds, T. Quatieri, R. Dunn. Speaker verification using adapted Gaussian Mixture Models. Digital Signal Processing, 2000, Vol. 10, pp: 19-41

[3] W. M. Campbell, D. E. Sturim, D. A. Reynolds and A. Solomonoff. SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. ICASSP 2006, pp: 97-100

[4] S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. on ASSP. 1980, Vol. 28, pp: 357-366

[5] NIST Speaker Recognition Evaluation Plan, Online Available http://www.nist.gov/speech/tests/sre/

[6] R. Vogt, S. Sridharan and Michael Mason. Making confident speaker verification decisions with minimal speech. IEEE Trans on ASLP 2010 vol. 18, no.6

[7] R. Vogt, C. Lustri, and S. Sridharan. Factor analysis subspace estimation for speaker verification with short utterances. Interspeech 2008

[8] Soonil Kwon, Shrikanth Narayanan. Robust speaker identification based on selective use of feature vectors. Pattern Recognition Letters. 2007, 28:85-89

[9] M. Nosratighods, Eliathamby Ambikairajah, Julien Epps and Michael John Carey. A segment selection technique for speaker verification. Speech Communication, 2010: 753~761

[10] A. Malegaonkar, A. Ariyaeeinia. On the enhancement of speaker identification accuracy using weighted bilateral scoring. ICCST 2008

[11] L. R. Rabiner, A tutorial on hidden Markov models and selected applicationsin speech recognition. Proc. IEEE 77 (1989), pp. 257–286.

[12] 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, 1998

[13] Linlin Wang and Thomas Fang Zheng. Creation of Time-Varying Voiceprint Database. Technical Session-6(Oral), Oriental-COCOSDA, 2010. Nov. 24-25,

[14] Wang D, Zhu X. Y., Liu Y.. Multi-Layer Channel Normalization for Frequency-Dynamic Feature Extraction. Journal of Software, v 12, n 9, September, 2005, p1523-1529

[15] 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, 1994, pp: 291-298