A K-Phoneme-Class based Multi-Model Method for Short Utterance Speaker Recognition

Chenhao Zhang[1], Xiaojun Wu[1], Thomas Fang Zheng[1*], Linlin Wang[1] and Cong Yin[1,2]
[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, China
[2]Taiyuan University of Technology
E-mail: {zhangchh, xjwu, wangll, yinc}@cslt.riit.tsinghua.edu.cn
* Corresponding Author: fzheng@tsinghua.edu.cn Tel/Fax: +86-10-62796393

***Abstract*—For GMM-UBM based text-independent speaker recognition, the performance decreases significantly when the test speech is too short. Considering that the use of text information is helpful, a K-phoneme-class scoring based multiple phoneme class speaker model method (shortened as K-phoneme-class based multi-model method, abbreviated as KPCMMM) is proposed including a phoneme class speech recognition stage and a phoneme class dependent multi-model speaker recognition stage, where K means the number of most likely phoneme classes to be used in the second stage. Two different phoneme class definitions, expert-knowledge based and data-driven, are compared, and the performance as a function of *K* is also studied. Experimental results show that the data-driven phoneme class definition outperforms the expert-knowledge based one, and that an appropriate *K* value can lead to much better performance. Compared with the baseline GMM-UBM system, the proposed KPCMMM can achieve a relative equal error rate (EER) reduction of 38.60% for text-independent speaker recognition with a length of less than 2 seconds of test speech.**

I. INTRODUCTION

Speaker recognition [1], aiming to automatically recognize the speaker identities, is becoming more and more attractive nowadays. It can be used in a wide range of applications including access control, providing forensic evidence, and user authentication in telephone banking, etc. Current speaker recognition technologies provide a satisfying performance when data is sufficient. However, in some situations, only a short utterance such as one or two words is available to recognize the speaker, and in other situations short utterances can provide a better user experience. In all such cases, the current technologies are unsatisfactory. In this paper, we focus on developing a method for short utterance speaker recognition (SUSR) where the test utterance contains only about 2 seconds' valid speech.

GMM-UBM [2] and GMM-SVM [3] are two popular speaker recognition technologies. In systems based on such structures, [6] illustrates the performance change with different valid test speech lengths on the NIST SRE 2005 [5] database, and it can be seen that the Equal Error Rate (EER) [5] increases sharply from 6.34% to 23.89% when the test speech is shortened from 20 seconds to 2 seconds. Furthermore, if the length is less than 2 seconds, the EER rises to as high as 35.00%.

In order to improve the performance of SUSR systems, some approaches have been proposed. The factor analysis subspace estimation introduced in [7] decreases the number of redundant model parameters to develop dominant speaker models. Some methods try to improve the performance by selecting segments with higher discriminability on speaker characteristics to perform speaker recognition [8]. The weighted bilateral scoring method is used to enhance the performance of speaker recognition in the scoring domain [9]. However, most of these above approaches show improvements with length among 5~10 seconds. There are still challenges when the speech is shorter.

It is no doubt that the performance of text-dependent (TD) speaker recognition is much better than that of text-independent (TI) speaker recognition when the length of speech is very short, because in this case the test data can better match the training data than in the TI case. This suggests us an idea to convert the TI SUSR into TD SUSR by integrating the speech recognition technology. In [10] it has showed that this method can help improve the text-independent speaker recognition.

Let us first introduce an idea of phoneme specific multi-model method for SUSR. During the training procedure, speech recognition is performed to generate a phoneme sequence. All data related to a certain phoneme will be collected together to train a speaker model specific to this phoneme and this speaker. During recognition, given an utterance, speech recognition will be first performed as in the training procedure to generate a phoneme sequence, each phoneme of which will be scored against phoneme specific speaker models of a speaker. After all phonemes have been scored, the score of the utterance against speakers will be obtained. This obviously can change a TI task into a TD one.

A question rises. None of the current speech recognizers is 100% correct, which will introduce errors in both the training and the recognition procedures. Error accumulations will lead to perhaps a much bigger performance decrease. Considering that speech recognition is not our final goal, instead we just need it to get the content information so as to convert a TI task into a TD one, we further propose to perform phoneme class recognition instead of phoneme recognition. The number of phoneme classes will be much smaller than that of phonemes and therefore the recognition errors will be much fewer provided that the phonemes are

well categorized. Either the expert knowledge based method or the data-driven one can be used for this purpose.

In practice, even a good performance phoneme class recognizer inevitably produces errors. To further eliminate the negative effect of these recognition errors, we propose not to use the top-1 phoneme class recognition results to perform phoneme class based multi-model speaker recognition, but during the phone class recognition procedure we will select top-*K* results for further speaker model scoring, which is referred to as top-*K* scoring in this paper. This only happens in the speaker recognition procedure yet not in the training procedure.

The above-described method is named as K-phoneme-class based multi-model method (KPCMMM). Depending on different phoneme clustering methods, it can be either expert-knowledge based or data-driven.

This paper is organized as follows. In Section II, the proposed KPCMMM framework for SUSR is detailed. In Section III, experimental results and analysis are given. Conclusions and future work are presented in Section IV.

II. THE K-PHONEME-CLASS BASED MULTI-MODEL SUSR FRAMEWORK

The proposed KPCMMM framework for SUSR can be shown in Fig. 1. There are three key parts: phoneme class definition, phoneme class dependent speaker model training, and K-phoneme-class based scoring.

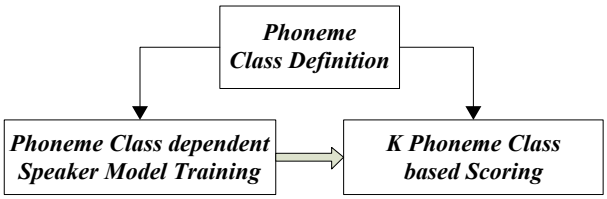

Fig. 1. The K-Phoneme-Class based Multi-Model SUSR Framework

A. Phoneme Class Definition

As mentioned above, an expert-knowledge based definition and a data-driven one will be addressed for phoneme class definition in this section,

The expert-knowledge based definition uses prior phonetics knowledge summarized by experts to categorize the phonemes. In this paper, the phoneme class definition in [11] is used and compared, which utilizes the Height and Backness information of the phonemes.

Given sufficient data, the data-driven phoneme class definition can be more performance oriented. We develop a vector quantization (VQ) phoneme clustering method, containing the following steps:

(1) Train a UBM with sufficient speech data. The data are chosen to cover all possible phonemes, and is also balanced according to concerned factors, such as channel and gender.

(2) Let $\{P_1, P_2, \cdots, P_N\}$ denote the entire phoneme set in one or several target languages, where N is the total number of possible phonemes. Collect data for each phoneme P_n and use the Maximum *a posteriori* (MAP) algorithm to generate the phoneme GMM model as in Equation (1).

$$p(P_i) = \sum_{m=1}^{M} w_m g(\mu_{im}, \sigma_{im}) = \sum_{m=1}^{M} w_m g_{im} \quad (1)$$

where M is the mixture number of the GMM model.

(3) Select initial J cluster centers (phoneme GMMs) from the phoneme GMM model set with the max-min criterion.

(4) The K-means algorithm [13] is used to cluster N phoneme models into J classes, where Kullback-Leibler (KL) divergence [14] is chosen as the distortion measure between Gaussian mixtures as described in Equations (2) and (3).

$$d(P_a, P_b) = \sum_{m=1}^{M} w_m KL(g_{am}, g_{bm}) \quad (2)$$

$$KL(g_{am}, g_{bm}) = \sum_{D} \frac{(\sigma_{am}{}^2 - \sigma_{bm}{}^2)^2 + (\mu_{am} - \mu_{bm})^2(\sigma_{am}{}^2 + \sigma_{bm}{}^2)}{\sigma_{am}{}^2 \sigma_{bm}{}^2} \quad (3)$$

The phoneme classes are defined as:

$$PC_j: P_{j,1}, P_{j,2}, \cdots, P_{j,n_j}, j \in [1, J], \sum_{j=1}^{J} n_j = N$$

where PC_j denotes the j-th phoneme class, and totally there are J phoneme classes. For PC_j, it contains n_j phonemes, and $P_{j,t}$ denote the t-th ($t \in [1, n_j]$) phoneme in PC_j.

To tell which of these two different phoneme class definitions is better, F-ratio, defined as the ratio of within-class variance to the between-class variance, is taken as the criterion on the phoneme clustering level. Equations (4) and (5) show how F-ratio is calculated:

$$F\text{-}ratio = \frac{\sum_{j=1}^{J} \frac{1}{n_j} \sum_{l=1}^{n_j} (\boldsymbol{\mu_{jl}} - \boldsymbol{\mu_j})(\boldsymbol{\mu_{jl}} - \boldsymbol{\mu_j})^T}{\sum_{j=1}^{J} \frac{n_j}{N} (\boldsymbol{\mu_j} - \boldsymbol{\mu})(\boldsymbol{\mu_j} - \boldsymbol{\mu})^T} \quad (4)$$

$$\boldsymbol{\mu_j} = \frac{1}{n_j} \sum_{t=1}^{n_j} \boldsymbol{\mu_{jt}}, \quad \boldsymbol{\mu} = \frac{1}{J} \sum_{j=1}^{J} \boldsymbol{\mu_j} \quad (5)$$

where $\boldsymbol{\mu_{jt}}$ represents the mean vector of the phoneme model $P_{j,t}$. The F-ratio curve of the definition derived through the data-driven method, as well as the F-ratio value of the expert-knowledge based definition, is shown in Fig. 2.

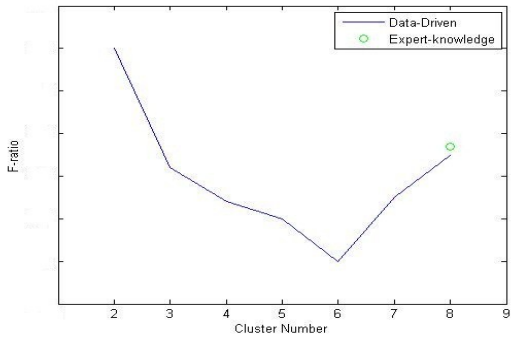

Fig. 2. F-ratio curve of the data-driven and expert-knowledge based definitions

It can be concluded that when the phoneme class number is 6, the data-driven definition achieves the minimum F-ratio value, which is better than the expert-knowledge based one.

B. Phoneme Class Dependent Speaker Model Training

Phoneme class dependent speaker models are trained as:

(1) Phoneme class UBM training. Phoneme recognition is performed on all utterances in the development set, and the

resulted phoneme sequences are divided and grouped into J phoneme classes. For phoneme class PC_j ($1 \le j \le J$), the phoneme class dependent UBM, denoted by λ_{ubm_j} is trained using the expectation-maximization (EM) algorithm.

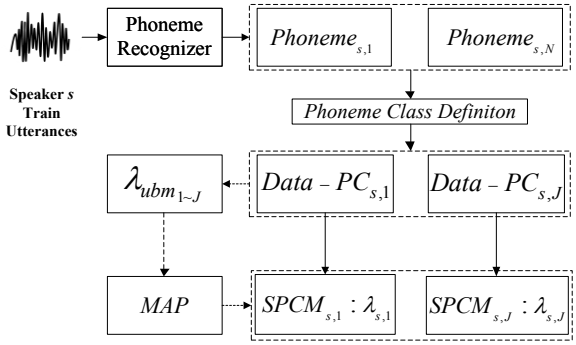

Fig. 3. Phoneme Class Dependent Speaker Model Training for Speaker s

(2) Phoneme class dependent speaker model training. For target speaker s, J phoneme class dependent models are trained, denoted by $\{\lambda_{s,j} : 1 \le j \le J\}$. For speaker s and phoneme class j, all data in $PC_{s,j}$ is used to generate $\lambda_{s,j}$ by adapting from λ_{ubm_j} with MAP algorithm as illustrated in Fig. 3.

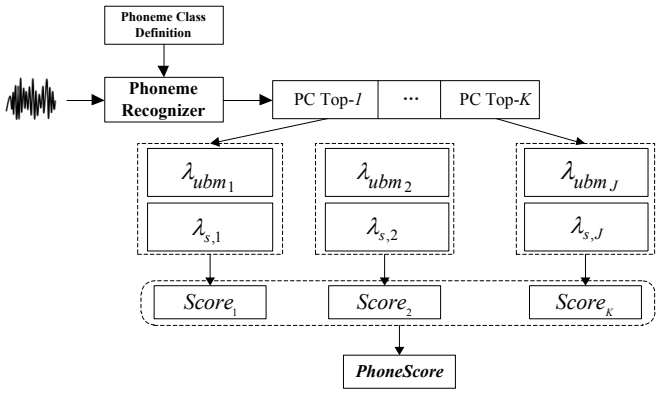

Fig. 4. Top-K Phoneme Class Scoring Illustration for One Phoneme of the Test Speech

C. K-Phoneme-Class based Scoring

In order to eliminate the influence of errors introduced by speech recognition, the K-NN algorithm is adopted, referred to as top-K scoring in this paper. During the phoneme class recognition stage of speaker recognition, the top-K phoneme class candidates of each phoneme in the utterance are taken instead of only top-1 as usual. The scores of this phoneme utterance against corresponding K phoneme class models are calculated, and fused as the final score of this phoneme. This is illustrated in Fig. 4 and in Equation (6).

$$P\text{-}Score_c = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{f_l}\sum_{t=1}^{f_l}\left[\log p\left(x|\lambda_{m,PC_k}\right) - \log p\left(x|\lambda_{ubm_{PC_k}}\right)\right] \quad (6)$$

where for each phoneme utterance in f_l frames, its corresponding top-K phoneme classes are PC_k ($1 \le k \le K$). The fusion method here is a simple weighted sum; that is to say, the final score is the weighted average score of the phonemes in test speech utterance as defined in Equation (7),

$$FusionScore_m = \frac{1}{\sum_{l=1}^{L} f_l}\sum_{l=1}^{L} f_l \, P\text{-}Score_l \quad (7)$$

where L is the number of phonemes contained in the test utterance.

III. Experimental Results and Analysis

A. Database and Setting up

The experiments were performed on a short utterance database specially created for SUSR, called SUD12. The SUD12 database consisted of 163 Chinese sentences each uttered by 57 Chinese speakers (29 males and 28 females). Speech was recorded in clean environments using microphone at 16 kHz sampling rate with 16-bit precision. The set of 163 sentences were grouped into two parts for training and testing purposes respectively. In the training part, there were 100 long sentences covering all Chinese vowels, balanced in term of phoneme so that all phoneme classes related models could be well trained. The testing part consisted of 63 short sentences, with average length of 2 in Chinese syllable, or less than 2 seconds at a normal reading speed. The distribution of lengths of utterances in the testing part is listed in Table I.

TABLE I
THE DISTRIBUTION OF THE LENGTH OF TEST UTTERANCES

Length (second)	Number of Sentences	Percent (%)
less than 0.5	38	60.3
0.5 to 1.0	15	23.8
1.0 to 2.0	10	15.9

The Chinese phoneme recognizer used here was trained using 50 hours' SONY Chinese speech [15]. The traditional MFCC features (12-dimensional MFCC coefficients, and their acceleration coefficients, delta coefficients, energy and zero static coefficients) were used. The recognizer was left-to-right no-skip HMM based.

There were two baseline systems, a speaker recognition system based on the conventional GMM-UBM and one using KPCMMM with the expert-knowledge based phoneme class definition [12]. The UBM consisted of 1,024 mixtures. The training data for UBM and the adaptation data for phoneme GMM models were taken from 863 CSL Corpus [16].

Features used for speaker recognition were also MFCCs, the only difference was that for speaker recognition they were 16-dimentional MFCC coefficients and their first derivative.

B. Results and Analysis

The EER and the minimum Detection Cost Function (minDCF) were used to evaluate the performance of speaker recognition systems, where the parameters of minDCF were taken as the same as in [5].

The EER curve of the SUSR system using the proposed KPCMMM with data-driven phoneme class definition as a function of K is shown in Fig. 5. It can be found that the proposed method achieved the best performance when K was 3, where the EER was 23.21%. Considering that the choosing of K value mainly depends on the performance of the phoneme class recognizer, the same K value was used in experiments with expert-knowledge based method. Results of

this method and the two baseline systems are listed in Table II and Fig. 5.

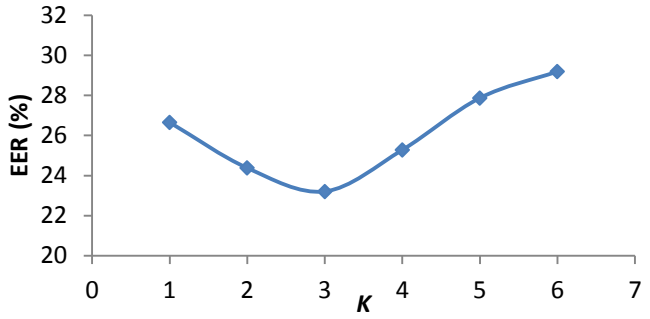

Fig. 5. EER curve of the K-phoneme-class based multi-model SUSR system (data-driven phoneme class definition) with different values of *K*

TABLE II
SPEAKER RECOGNITION PERFORMANCE COMPARISION

Method	EER (%)	minDCF (10^{-2})
GMM-UBM	37.81	10.52
Expert-KPCMMM	27.73	9.72
Data-Driven-KPCMMM	23.21	9.22

Compared with the traditional GMM-UBM method, the proposed method with different phoneme class definitions achieved relative EER reductions of 26.64% and 38.60% respectively. These results show that the KPCMMM can emphasize the match between the models and the test utterances.

For the proposed KPCMMM, the data-driven phoneme class definition outperforms the expert-knowledge based one, with a relative EER reduction of 16.30%. The reason is believed to lie in that the data-driven method is performance oriented.

It is also found that using top-K phoneme recognition results for further multi-modal speaker recognition will be much better than just using the top-1 results, which can avoid the error accumulation introduced by the speech recognition stage. The value of *K* should be appropriate; being too small (as 1) or too big (using all classes) does not work well.

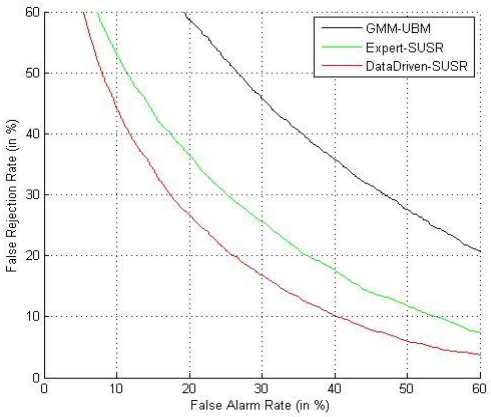

Fig. 6. DET Curve Comparison among GMM-UBM, Expert-Knowledge based KPCMMM, Data-Driven KPCMMM

IV. CONCLUSIONS AND FUTURE WORK

In this paper, we propose a text-independent short utterance speaker recognition method integrating the phoneme class definition and the K-phoneme-class based multi-model method. The experimental results show that the proposed KPCMMM with data-driven phoneme class definition can achieve a better result for short test utterances.

In this paper there is an assumption that the training data is sufficient for phoneme class models. By setting up a set of reference speakers with sufficient training data, this assumption is not necessary. This is one of our future research focuses. Combining the expert knowledge and the data-driven method is another focus in future research.

REFERENCES

[1] J. P. Campbell. Speaker recognition: a tutorial. Proceedings of the IEEE, 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 et al. 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, pp. 1182-1192

[7] P. Kenny, G. Boulianne, and P. Dumouchel. Eigenvoice modeling with sparse training data. IEEE Trans. on Speech and Audio Processing, 2005, vol. 13, no. 3, pp. 345-354

[8] M. Nosratighods, E. Ambikairajah, J. Epps and M. J. Carey. A segment selection technique for speaker verification. Speech Communication, 2010, pp. 753-761

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

[10] Yu Tsao et al. An acoustic segment model approach to incorporating temporal information into speaker modeling for text-independent speaker recognition. ICASSP,2010, pp. 4422-4425

[11] N. Fatima, X.-J. Wu, T. F. Zheng and G. Wang. A universal phoneme-set based language independent short utterance speaker recognition. NCMMSC, 2011

[12] C.-H. Zhang, L.-L. Wang, J. R. Jang and T. F. Zheng. A multi-model method for short-utterance speaker recognition. APSIPA 2011

[13] A. V. Hall. Methods for demonstrating resemblance in taxonomy and ecology, Nature, 1967, vol. 214, pp. 830-831,

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

[15] L.-Q. Liu, T. F. Zheng, and W.-H. Wu, english alphabet recognition based on chinese acoustic modeling, COLIPS 2006, pp.463-472

[16] D. Wang, X. Zhu, Y. Liu. Multi-layer channel normalization for frequency-dynamic feature extraction. Journal of Software, 2005, vol. 12, no. 9, pp.1523-1529