Time-Frequency Cepstral Features and Combining Discriminative Training for Phonotactic Language Recognition

Yan Deng, Wei-Qiang Zhang, Yan-Min Qian and Jia Liu
Tsinghua National Laboratory for Information Science and Technology
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
Email: y-deng05@mails.thu.edu.cn, wqzhang@tsinghua.edu.cn, qianym07@mails.thu.edu.cn, liuj@tsinghua.edu.cn

***Abstract*—The performance of the phonotactic system for language recognition depends on the quality of the phone recognizers. To improve the performance of the recognizers, this paper investigates the use of new acoustic features and discriminative training techniques for phone recognizers. The commonly used features are static ceptral coefficients appended with their first and second order deltas. This configuration may be not optimal for phone recognition in phonotactic language recognition systems. In this paper, a time-frequency cepstral (TFC) feature is proposed based on our previous work in acoustic language recognition systems. The feature is extracted as follows: first a temporal discrete cosine transform (DCT) is carried out on the cepstrum matrix, and then select the transformed elements in a specific area using the variance maximization criterion. Different parameters are tested to obtain the optimal configuration. Also, we adopt the feature minimum phone error (fMPE) method for discriminative training of phone models to obtain better phone recognition results for further improvement. The effectiveness of the two techniques is demonstrated on the NIST Language Recognition Evaluation (LRE) 2007database, including the 30 second, 10 second and 3 second closed-set test conditions.**

***Index Terms*—phonotactic language recognition, phone recognizer, time-frequency cepstrum (TFC), feature minimum phone error (fMPE)**

I. Introduction

Language recognition is to determine the language identity of a given speech segment. There are two types of commonly used systems for language recognition: acoustic systems and phonotactic systems. The acoustic systems use Gaussian Mixture Models (GMM) [1] or support vector machines (SVM) [2] to model the long term spectral characteristics of speech, also referred to as the spectral systems. While the phonotactic systems use N-gram language model [3], binary tree [4] or SVM [5] to model the language dependent lexical constraints of token sequences or lattices decoded from speech. For utilizing the intermediate results of recognizers (or tokenizers), they are also referred to as the token-based systems. Many state-of-the-art systems include both techniques to achieve optimal performance.

This paper will focus on the phonotactic system called parallel phone recognizers followed by vector space modeling (PPRVSM) [5]. In this system, multiple language dependent phone recognizers are used to map speech segments spoken in any language to phone sequences or lattices [6]. Usually, the mapping is performed without any phonotactic constraints. Phone sequences or lattices are then used to estimate phonetic N-gram statistics which will be modeled by SVM classifiers. Generally, the performance of the PPRVSM system depends on the performance and consistency of the front-end phone recognizers. Better phone recognition leads to better N-gram estimates, which in turn leads to better language recognition results. Several methods have been proposed for the phone recognizers, such as using the Artificial Neural Network-Hidden Markov Model (ANN/HMM) architecture instead of GMM/HMM to improve the performance of phone recognizers [7], substituting the context-dependent (CD) HMM phone models for the context-independent (CI) ones to remove the context variability and using model adaption to remove the channel variability [8]. Also, there are other ways to improve the phone recognizers.

Usually, the Mel-frequency cepstral coefficients (MFCC) and perceptual linear prediction (PLP) features are used for phone recognition [5, 6]. To utilize the long temporal information of time-varying spectral features and reduce the effect of channel convolution distortions, the first and second order deltas are calculated and appended to the basic ones. However, this configuration may be not optimal for phone recognition. Firstly, the deltas can be obtained by taking high-pass linear filtering of the basic ones. According to the information theory, it may be redundant. Secondly, the first and second order deltas may bring correlation into the concatenated feature, which will depress the performance of the commonly used diagonal GMM. Therefore, the simple concatenation may not be used to select the most informative elements. In our previous work, we have presented a time-frequency cepstrum (TFC) feature for the GMM-based acoustic system [9], which is obtained by performing a

Manuscript received July 7, 2010; revised August 13, 2010; accepted August 15, 2010.

doi:10.4304/jcp.6.2.178-183

temporal discrete cosine transform (DCT) on the cepstrum matrix. Experimental results have shown that it outperforms the widely used shifted delta cesptrum (SDC) feature [9]. In this paper, we will extend this work for phone recognition in the PPRVSM system. The discriminative training method is also used for further improvement.

This paper is organized as follows. Section 2 gives a brief description of the PPRVSM system: the phonotactic feature extraction and vector space modeling (VSM). The TFC feature extraction and discriminative training algorithm are presented in section 3 and 4 respectively. The experimental results on the NIST LRE 2007 evaluation database will be shown in section 5 followed by a conclusion in section 6.

II. THE PPRVSM SYSTEM

A. Phonotactic Feature Extraction

In the PPRVSM language recognition, several parallel phone recognizers (PPR) are used to decode the speech into phone sequences or lattices for following analysis. We developed a Mandarin phone recognizer using the GMM/HMM architecture and trained on the conversational telephone data including about 30 hours of speech. There are 64 phone models for the phone recognizer. Each phone model is a tied-state left-to-right CD HMM with 32 Gaussians per state. Compared with CI phone models, the CD models have the advantage that it can handle the context variability [8].

Typically, the phonetic N-gram probabilities can be estimated on either phone sequences or lattices. Using phone lattices often produces better language recognition results than sequences [6]. In this paper, lattices generated from the phone recognizers are used for phonotactic feature extraction. A phone lattice is a rich and compact representation of multiple hypotheses with acoustic likelihoods, from which the expected counts of phonetic N-grams are estimated. The expected counts of phonetic N-grams can be understood as an extension of standard counts. Given a hypothesized phone sequence: $S = s_1 ... s_i s_{i+1} ... s_n$, the phonetic N-grams are created by grouping N phones at a time to form, such as $s_i ... s_{i+N-1}$. The count of the N-gram $s_i ... s_{i+N-1}$ is the number of occurrence of $s_i ... s_{i+N-1}$ in the sequence S. To extend to the lattice $\mathcal{L}$, the expected counts are calculated over all possible hypotheses in the lattice [6]:

$$c(s_i ... s_{i+N-1} \mid \mathcal{L}) = \sum_{S \in \ell} p(S \mid \mathcal{L}) c(s_i ... s_{i+N-1} \mid S) = \sum_{s_i ... s_{i+N-1} \in \mathcal{L}} [\alpha(s_i) \beta(s_{i+N-1}) \prod_{j=i}^{i+N-1} \xi(s_j)] .$$

where, $p(S \mid \mathcal{L})$ is the probability of the sequence S in the lattice $\mathcal{L}$, $\alpha(s_i)$ is the forward probability of the starting node in the N-gram $s_i ... s_{i+N-1}$, $\beta(s_{i+N-1})$ is the backward probability of the ending node, $\xi(s_j)$ is the posterior probability of the edge s_j. The probability of the N-gram $s_i ... s_{i+N-1}$ in the lattice is then computed as follows:

$$p(s_i ... s_{i+N-1} \mid \mathcal{L}) = \frac{c(s_i ... s_{i+N-1} \mid \mathcal{L})}{\sum_i c(s_i ... s_{i+N-1} \mid \mathcal{L})} .$$

B. Vector Space Modeling (VSM)

Recently, VSM has become a popular technique in phonotactic language recognition systems [5]. In the VSM, each speech segment is represented by a super-vector and then modeled using SVM.

The key point of VSM is the sequence kernel construction. We adopt the term frequency log-likelihood ratio (TFLLR) kernel that has been proven to be useful for phonetic speaker recognition [10]. The TFLLR kernel is constructed by scaling the phonetic N-gram probabilities that are estimated on a given phone lattice. In general, the kernel function between two lattices can be represented as follows:

$$K(\mathcal{L}_1, \mathcal{L}_2) = \sum_i p_n(s_i ... s_{i+N-1} \mid \mathcal{L}_1) * p_n(s_i ... s_{i+N-1} \mid \mathcal{L}_2) = \sum_i \frac{p(s_i ... s_{i+N-1} \mid \mathcal{L}_1)}{\sqrt{p(s_i ... s_{i+N-1} \mid all)}} * \frac{p(s_i ... s_{i+N-1} \mid \mathcal{L}_2)}{\sqrt{p(s_i ... s_{i+N-1} \mid all)}} \quad (1)$$

The denominator $p(s_i ... s_{i+N-1} \mid all)$ is the probability of $s_i ... s_{i+N-1}$ in all the phone lattices in the training set. This insures that the kernel function is not dominated by N-grams with large probabilities. The inner product in Eq. (1) indicates that if the same N-grams are present in the two lattices, then there will a high degree of similarity between them two, and vice versa.

The back-end of VSM is the SVM training and scoring. The training is carried out with a "one-versus-rest" strategy. The samples in the target language are collected as the positive set and the remaining in other classes as the negative one, then carry out training between them two.

Given the vector $\vec{X} = \{p_n(s_1 \mid \mathcal{L}), ..., p_n(s_i ... s_{i+N-1} \mid \mathcal{L})\}$ and kernel function $K(X, X_l)$, the SVM scoring function is:

$$f(X) = \sum_l \alpha_l t_l K(X, X_l) + d . \quad (2)$$

A decision is based on the output of Eq. (2) compared to a predefined threshold. The X_l are support vectors obtained under the Mercer condition.

Compared with low-order N-grams, high-order ones are usually more discriminative for language recognition. But there is a problem that the number of N-grams grows exponentially as the order N increases, which results in a high dimensional phonotactic feature vector. It is a challenge for SVM training. However, researchers have illustrated that not all the N-grams are necessary in building the VSM [11]. So it is important to pick out the most discriminative phonetic N-grams for language recognition. A selection method proposed in our previous

work is adopted in this paper [12]. The approach can be decomposed into two stages: selection and construction. The selection is the process of picking out the most discriminative low-order N-grams using the maximum mutual information criterion. While the construction is designed to create high-order N-grams based on the selected low-order ones and the phone set.

III. TFC FEATURE EXTRACTION

The TFC feature is initially proposed in the GMM-based acoustic system for language recognition [9]. The extraction is performed as follows: several successive frames of basic features within a context width are extracted first to form a cepstrum matrix. Then a DCT is implemented on the cepstrum matrix in the temporal direction to remove the correlation. Finally, the elements in the upper-left triangular area are selected by scanning in a zigzag order.

There are two benefits by extracting features in such a way. Firstly, the procedure of TFC feature extraction is equivalent to performing a two dimensional (2D) DCT on the spectrum-time matrix. The 2D DCT approach can be interpreted as a compression of the information by a DCT truncation. The truncation of the higher order vectors helps to reduce the variability caused by small scale acoustic events. Also, compared with the SDC feature, the elements can be selected with a greater variability for the TFC feature. Rather than simply appending the delta elements to the basic ones.

In the GMM-based acoustic systems, the normalized variances (normalized by the maximum elements) of the cepstrum matrix after a temporal DCT is nearly a triangle, so we can perform a zigzag scan to select elements in the upper-left area to obtain the TFC feature. But for phone recognition, the optimal configuration will not be different from that of the GMM system. Usually, the context width is about 20 frames in the GMM system while it becomes 9 for phone recognition, which is much smaller. Then the variances pattern will become much narrower accordingly. To give a simple illustration, we plot the normalized variance of each element in the cepstrum matrix after a temporal DCT in Fig. 1. In this figure, the matrix is created using successive 9 frames of 20-dimensional MFCC basic feature vector and the variance was computed on the data corpus used for training phone models. Fig. 1 shows that there are obviously other possible configurations besides the triangle adopted in the GMM system. Maybe a rectangle will be better for phone recognition.

IV. DISCRIMINATIVE TRAINING OF THE PHONE MODELS

In acoustic language recognition, Maximum Mutual Information (MMI) has been proposed in the GMM system and great improvements can be obtained compared with the Maximum Likelihood (ML) training [13]. In phonotactic systems, the discriminative training has also been adopted in the language modeling such as using SVM [5]. Few researchers pay attention to the discriminative training of phone models, which will result in better phone sequences or lattices.

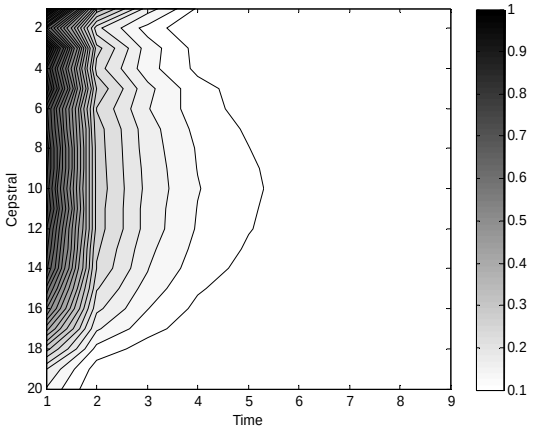

Figure 1. The normalized variances of each element in the cepstrum matrix after a horizontal DCT.

Discriminative training of acoustic models has been widely used in speech recognition to improve the recognition accuracy, such as MMI, Minimum Classification Error (MCE), Minimum Phone Error (MPE) [14] and fMPE [15]. The techniques are all introduced for discriminative training of HMM parameters with their corresponding objective function. In application, all techniques are performed on acoustic models except fMPE, which applies to the acoustic features. This makes possible things that are impossible with other discriminative training techniques which are done directly on acoustic models, such as building a new system using the new features. In this paper, we will adopt the fMPE technique for discriminative training of phone models used in phonotactic language recognition.

fMPE is a previously introduced form of discriminative training using the same objective function as MPE, in which offsets to the features are obtained by training a projection from a high-dimension feature space based on posteriors of Gaussians [15]. fMPE is performed by transforming the acoustic feature with a kernel-like method. Let x_t be the original features and y_t be the transformed features. The formula is represented in the following:

$$y_t = x_t + \boldsymbol{M} h_t .$$

where, $\boldsymbol{M}$ is the transform matrix that needs to be estimated by optimizing the MPE objective function and h_t is the expanded high dimensional feature derived based on posteriors of Gaussians. The calculation of h_t and $\boldsymbol{M}$ can be found in [15].

For fMPE training, we first need to generate lattices by decoding the training data with a weak language model [16]. The lattices are then used to produce the MPE statistics. Also a GMM is needed to obtain Gaussian posteriors. In this paper, we use 1000 Gaussians to calculate the offset features within a context width of 5.

In our experiments, we often run 3-4 iterations to obtain an optimal result.

V. EXPERIMENTS

A. Experimental Setup

The experiments are carried out on the NIST 2007 Language Recognition Evaluation (LRE) test set under the closed condition. There are 7530 utterances in total, spanning the 3, 10 and 30 second conditions. The task is to recognize 14 target languages: Arabic, Bengali, Chinese, English, Farsi, German, Hindustani, Japanese, Korean, Russian, Spanish, Tamil, Thai and Vietnamese. The training data we use for VSM includes Callfriend, the evaluation and development data provided by NIST in the previous LRE. As the speech is relatively long, the voice activity detection is used to segment each utterance into segments with about 30 seconds of speech in length.

The configuration of the Mandarin phone recognizer is given in section 2.1. For acoustic feature extraction, standard 12 MFCC coefficients with c0 are extracted every 10 ms over a 25 ms hamming window. These features are augmented by their first and second order deltas, resulting in a 39 dimension feature vector. Also, cepstral mean subtraction and variance normalization are both applied to remove the channel variability.

For the phonotactic feature selection, the top 20% of the low-order N-grams are selected based on the mutual information value, which has been defined according to our previous work [12]. The SVM training requires pre-computing all of the inner products between the data points. This approach is good for feature expansion to high dimension when the number of data points is not too large.

B. Experimental Results

We demonstrate the effectiveness of our approaches under 3, 10 and 30 second conditions. Both detection cost function (DCF, Cavg*100) and equal error rate (EER) are utilized to summarize the results.

TABLE I. COMPARISON OF TFC WITH DIFFERENT RECTANGULAR AREA

EER Cavg*100	30 second	10 second	3 second
MFCC 39	2.95 2.68	9.25 8.88	20.73 20.44
TFC 13×2+13	3.09 2.83	9.67 9.38	21.14 20.53
TFC 13×3+13	2.60 2.43	8.61 8.22	20.10 19.73
TFC 13×4+13	2.54 2.45	8.58 8.34	20.74 20.10

The first experiment is to find the optimal configuration of the reserved area. In order to give fair comparison, we fix the context width to 9. In this work, we adopt the rectangular shape according to Fig. 1 and test three TFC feature configurations where static cepstral coefficients are concatenated with the elements of a cepstrum matrix obtained by a temporal DCT. Settings referred to as TFC $N\times O$ defines a TFC feature where a temporal DCT of order O is performed on a context window of 9 frames successive N-dimension MFCCs (C_0 to C_{N-1}). The language recognition results are given in Table I. From this table, we can see that if we fix the dimension to 39, the performance is a little worse than the MFCC feature. However, when we increase the order of the temporal DCT (the width of the rectangular), then we will get the best performance using 52 dimensions (39+13 static parameters). Both the EER and Cavg*100 are lower than those of the MFCC feature under the three test conditions, with a relative decrease of 3%-7% in general. The improvements attained are comparable with those in the GMM system [9], which demonstrates that the proposed TFC feature is effective for the diagonal GMM classifiers employed in both the GMM and PPRVSM system.

Our second experiment is to show the effectiveness of the fMPE approach. The language recognition results of the PPRVSM system using fMPE trained phone recognizer are summarized in Table II and III. The original MFCC feature and the propose TFC feature are both used for experiments. We can see that fMPE training will produce better language recognition results than the maximum likelihood training method for both MFCC and TFC under the three test conditions. Although only slight improvements are obtained, it indicates that fMPE training will produce better phone recognizers, which in turn leads to better language recognition results. However, the improvements attained using fMPE instead of ML (Table II) is not as significant as using TFC instead of MFCC (Table I). This is because the fMPE is implemented by just adding offsets to the original features to increase the discriminative ability of phone models, the amount of information will not increase. But the TFC feature is a novel one extracted in a completely different way from the MFCC feature, by which more information can be utilized for phone recognition. As more useful information are extracted to improve the phone recognition results, then the fMPE approach will not be as useful for the TFC feature as for the MFCC feature, which can be seen from Table II and III.

TABLE II. COMPARISON OF DIFFERENT TRAINING METHODS FOR PHONE MODELS (MFCC 39)

EER Cavg*100	30 second	10 second	3 second
ML	2.95 2.68	9.25 8.88	20.73 20.44
fMPE	2.69 2.45	8.92 8.45	20.36 20.29

TABLE III. COMPARISON OF DIFFERENT TRAINING METHODS FOR PHONE MODELS (TFC 52)

EER Cavg*100	30 second	10 second	3 second
ML	2.60 2.43	8.61 8.22	20.10 19.73
fMPE	2.45 2.35	8.24 8.07	20.56 19.77

Finally, we compare the improved Mandarin phone recognizer with others, which are provided by the Faculty

of Information Technology of the Brno University of Technology [7]. They are all ANN/HMM hybrid systems and trained on the Czech, Hungarian and Russian separately. Table IV summarizes the language recognition results using different phone recognizers in the front-end. As expected, the results show that the original and improved Mandarin phone recognizers are all better than others under all test conditions. We also conduct experiments of fusing multiple phone recognizers. We adopt the "linear discriminant analysis (LDA) + Gaussian" method. First we combine the Czech, Hungarian and Russian, and then add Mandarin to the system for fusion. The improvements shown in Table IV indicate that the three Mandarin phone recognizers are complementary with others. It is because that the Czech, Hungarian and Russian phone recognizers use the ANN/HMM architecture while the Mandarin phone recognizer adopts the GMM/HMM architecture. We further improve the language recognition performance by fusing all systems together, as shown in the last row of Table IV.

TABLE IV. COMPARISON AND FUSION OF MULTIPLE PHONE RECOGNIZERS

EER Cavg*100	30 second	10 second	3 second
Mandarin (MFCC-39)	2.95 2.68	9.25 8.88	20.73 20.44
Mandarin (TFC-52)	2.60 2.43	8.61 8.22	20.10 19.73
Mandarin-fMPE (TFC-52-fMPE)	2.45 2.35	8.24 8.07	20.56 19.77
Czech (CZ)	4.59 4.25	11.60 11.08	23.27 23.64
Hungarian (HU)	3.54 3.32	9.96 10.00	22.17 22.18
Russian (RU)	4.30 4.04	11.68 10.78	22.62 21.73
CZ, HU&RU	2.59 2.29	7.41 7.63	19.07 17.09
CZ, HU, RU& MFCC- 39	1.64 1.45	5.58 5.45	16.03 15.81
CZ, HU, RU&TFC-52	1.46 1.37	5.26 5.21	15.73 15.16
CZ, HU, RU&TFC-52-fMPE	1.41 1.30	5.27 5.28	14.96 15.21
All	1.34 1.20	5.24 5.14	15.22 15.08

VI. CONCLUSIONS

We have presented two methods to improve the performance of the PPRVSM language recognition system. First, we proposed the TFC features for phone recognition, an alternative to MFCC that, in our advice, has more perceptual grounds, wider flexibility, and give better results than MFCC. Then, we adopt fMPE to further improve the quality of the HMM based phone recognizers. Results show that the baseline system benefits from these techniques. Applied together, the system using Mandarin phone recognizer gets a significant improvement in EER, with relative improvements of 16.95% for the 30 second test condition, 10.92% for the 10 second test condition and 3.04% for the 3 second test condition. Finally, the experiment for comparison and fusion of multiple phone recognizers indicates that the improved recognizer is comparable and complementary with other recognizers. The final EER of the fused PPRVSM system achieves 1.34%, 5.24% and 15.22% for the 30 second, 10 second and 3 second test set respectively.

As the results are obtained using the 52-dimension TFC feature, which is higher than MFCC. Then the computation will increase accordingly. For further research, a proper feature selection method should be developed to reduce the dimension to 39. And the selected 39-dimension feature has the same amount of information as the 52-dimension TFC feature for phone recognition.

ACKNOWLEDGMENT

This work is supported by National Natural Science Foundation of P. R. China and Microsoft Research Asia (60776800), National Natural Science Foundation of P. R. China and Research Grants Council (60931160443), National High Technology Research and Development Program of China (863 Program) (2008AA02Z414, 2008AA040201).

REFERENCES

[1] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller, "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features," *In Proc. ICSLP'02*, Denver, Colorado, USA, 2002, pp. 33–36.

[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 and Language*, vol. 20, pp. 210–229, 2006.

[3] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Trans. Speech and Audio Processing*, vol. 4, pp. 31–44, 1996.

[4] J. Navratil, "Recent advances in phonotactic language recognition using binary decision trees," *In Proc.ICSLP'06*, Pittsburgh, USA, 2006, pp. 421-424.

[5] H. Li, B. Ma, and C. H. Lee, "A vector space modeeling approach to spoken language identification," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, pp. 271–284, 2007.

[6] J. L. Gauvain, A. Messaoudi, and H. Schwenk, "Language recognition using phone lattices," *In Proc. ICSLP'04*, Jeju Island, Korea, 2004, pp. 1283–1286.

[7] P. Matejka, P. Shwarz, J. Cernocky, and P. Chytil, "Phonotactic language identification using high quality phoneme recognition," *In Proc. Interspeech'05*, Lisbon, Portugal, 2005, pp. 2237-2240.

[8] M. F. BenZeghiba, J. Gauvain, and L. Lamel, "Context-dependent phone models and models adaptation for phonotactic language recognition," *In Proc. Interspeech'08*, Brisbane, Australia, 2008, pp. 313-316.

[9] W.-Q. Zhang, L. He, Y. Deng, J. Liu, and M. T. Johnson, "Time-frequency cepstral features and heteroscedastic linear discriminant analysis for language recognition". *IEEE Trans. on Audio, Speech, and Language Processing*, in press.

[10] W. M. Campbell, J. P. Campbell, D. A. Reynolds, D. A. Jones, and T. R. Leek, "Phonetic speaker recognition with support vector machines," *Advances in Neural Information Processing System 16*, Sebastian Thrun, Lawrence Saul, and Bernhard Scholkopf, Eds. MIT Press, Cambrige, MA, 2004.
[11] F. S. Richardson, and W. M. Campbell, "Language recognition with discriminative keyword selection," *In Proc. ICASSP'08*, Las Vegas, Nevada, USA, 2008, pp. 4145-4148.
[12] Y. Deng, W. Q. Zhang, and J. Liu, "Language recognition based on discriminative vector space model," *Journal of Nanjing University of Science and Technology*, vol. 33, pp. 138-144, 2009.
[13] L. Burget, P. Matejka, and J. Cernocky, "Discriminative training techniques for acoustic language identification," *In Proc. ICASSP'06*, Toulouse, France, pp. 209-212, 2006.
[14] D. Povey, "Discriminative training for large vocabulary speech recognition," [Ph.D thesis]. Cambridge University, Cambridge, UK, 2004.
[15] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "fMPE: Discriminatively trained features for speech recognition," *In Proc. ICASSP'05*, Philadelphia, PA, USA, pp. 961-964, 2005.
[16] R. Schluter, B. Muller, F. Wessel, and H. Ney, "Interdependence of language model and discriminative training," *In Proc. ASRU Workshop*, Keystone, Colorado, USA, pp. 119–122, 1999.

Yan Deng was born in Hunan, China, in 1982. She received the B.S. degree in communication engineering from National University of Defense Technology, Changsha, China, in 2005.

She is currently a Ph.D. candidate in the Department of Electronic Engineering, Tsinghua University, Beijing, China. Her research focuses upon language recognition and speaker recognition.

Wei-Qiang Zhang was born in Hebei, China, in 1979. He received the B.S. degree in applied physics from University of Petroleum, Shangdong, China, in 2002, the M.S. degree in communication and information systems from Beijing Institute of Technology, Beijing, China, in 2005, and the Ph.D. degree in information and communication engineering from Tsinghua University, Beijing, in 2009.

He is currently a Research Assistant at the Department of Electronic Engineering, Tsinghua University. His research interests are in the area of speech and signal processing, primarily in parameter estimation, higher order statistics, time–frequency analysis, speaker recognition, and language recognition.

Yan-Min Qian was born in Zhejiang, China, in 1984. He received his B.S degree in the Department of Electronic and Information Engineering from Huazhong University of Science and Technology, China, in 2007.

He is currently a Ph.D. candidate in the Department of Electronic Engineering, Tsinghua University, China. His research focuses upon fast decoding, robust speech recognition and large vocabulary speech recognition.

Jia Liu was born in Fujian, China, in 1954. He received the B.S., M.S., and Ph.D. degrees in communication and electronic systems from Tsinghua University, Beijing, China, in 1983, 1986, and 1990, respectively.

He worked at the Remote Sensing Satellite Ground Station, Chinese Academy of Sciences, after the Ph.D. degree and worked as a Royal Society Visiting Scientist at the Cambridge University Engineering Department, Cambridge, U.K., from 1992 to 1994. He is now a Professor in the Department of Electronic Engineering, Tsinghua University. His research fields include speech recognition, speaker recognition, language recognition, expressive speech synthesis, speech coding, and spoken language understanding.