IEICE TRANS. INF. & SYST., VOL.E94–D, NO.3 MARCH 2011

PAPER

Language Recognition Based on Acoustic Diversified Phone Recognizers and Phonotactic Feature Fusion

Yan DENG[†a)], ***Student Member***, **Wei-Qiang ZHANG**[†], ***Nonmember***, **Yan-Min QIAN**[†], ***Student Member***, ***and* Jia LIU**[†], ***Nonmember***

SUMMARY One typical phonotactic system for language recognition is parallel phone recognition followed by vector space modeling (PPRVSM). In this system, various phone recognizers are applied in parallel and fused at the score level. Each phone recognizer is trained for a known language, which is assumed to extract complementary information for effective fusion. But this method is limited by the large amount of training samples for which word or phone level transcription is required. Also, score fusion is not the optimal method as fusion at the feature or model level will retain more information than at the score level. This paper presents a new strategy to build and fuse parallel phone recognizers (PPR). This is achieved by training multiple acoustic diversified phone recognizers and fusing at the feature level. The phone recognizers are trained on the same speech data but using different acoustic features and model training techniques. For the acoustic features, Mel-frequency cepstral coefficients (MFCC) and perceptual linear prediction (PLP) are both employed. In addition, a new time-frequency cepstrum (TFC) feature is proposed to extract complementary acoustic information. For the model training, we examine the use of the maximum likelihood and feature minimum phone error methods to train complementary acoustic models. In this study, we fuse phonotactic features of the acoustic diversified phone recognizers using a simple linear fusion method to build the PPRVSM system. A novel logistic regression optimized weighting (LROW) approach is introduced for fusion factor optimization. The experimental results show that fusion at the feature level is more effective than at the score level. And the proposed system is competitive with the traditional PPRVSM. Finally, the two systems are combined for further improvement. The best performing system reported in this paper achieves an equal error rate (EER) of 1.24%, 4.98% and 14.96% on the NIST 2007 LRE 30-second, 10-second and 3-second evaluation databases, respectively, for the closed-set test condition.

key words: *language recognition, parallel phone recognition followed by vector space modeling (PPRVSM), acoustic diversified phone recognizers, feature fusion, logistic regression optimized weighting (LROW), time-frequency cepstrum (TFC)*

1. Introduction

Language recognition is the process of determining the language identity from a sample of speech. It is an essential technology in many applications, such as multilingual speech recognition, spoken language translation and information security [1], [2]. Generally speaking, language recognition acts as a front-end for many multilingual speech processing systems, where the language identities of speech need to be established for further information extraction.

Language recognition can be performed using information from multiple resources. In the past few decades, researchers have developed many different systems to explore the discriminative information. Currently, two types of systems are widely used for language recognition: acoustic systems and phonotactic systems. Acoustic systems use Gaussian Mixture Models (GMM) [3] or support vector machines (SVM) [4] to model the long-term spectral characteristics. Phonotactic systems use parallel phone recognizers (PPR) to convert the speech into phone sequences or lattices and then perform phonotactic analysis using an N-gram language model [2], a binary tree [5] or a vector space model (VSM) [6]. State-of-the-art systems often include both techniques to achieve optimal performance. This paper focus on a phonotactic system called the parallel phone recognizer followed by vector space modeling (PPRVSM), which employs VSM to model the phonotactics of different languages [6]. In the system, phone lattice is used due to its superiority over phone sequence [7].

Generally, the PPRs of a phonotactic system can be developed in two ways. One is to train phone recognizers on multiple language-specific speech data with different phone sets to provide phonetic diversification [2]. The other is to train phone recognizers on the same language-specific speech data with one phone set but using different acoustic models to provide acoustic diversification [8]. The PPRs developed using phonetic diversification can make sure that the phone coverage will be sufficient enough to cover the sound units of all target languages in a language recognition task. But one problem is that word or phone-level transcription is needed to train a phone recognizer. While an increased number of phone recognizers provides better performance, it also requires more transcribed training data. For acoustic diversification, the phone recognizers are often well-designed to emphasize different acoustic aspects of the speech to achieve a good diversification. Compared with phonetic diversification, acoustic diversification can be adopted to overcome the difficulty of collecting large amounts of transcribed data. Moreover, comparable results can be achieved under both diversifications [8]. Therefore, the acoustic diversification becomes more attractive when building the PPR front-end.

However, there are three main problems for the acoustic diversified phone recognizers. First, the phone sequences or lattices are homogeneous since the same training data and phone set are used. It will be a great challenge to develop a set of phone recognizers to achieve a good acoustic diver-

---

Manuscript received August 30, 2010.
Manuscript revised November 8, 2010.

[†] The authors are with Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China.
a) E-mail: dengyan05@gmail.com

DOI: 10.1587/transinf.E94.D.679

sification. In reference [8], different model structures and training paradigms are employed to build multiple phone recognizers. But it's not the best way as the same acoustic feature is used. While different acoustic features extract different information from the speech, they will provide better acoustic diversification. Second, these phone recognizers are fused by combining multiple phone sequences to retain more information for fusion. It has been demonstrated that this method is more effective than score fusion [8]. But it becomes infeasible when phone lattices are used instead of sequences as the acoustic scores of different phone recognizers are compatible. Finally, as the same training data and phone set are used, high correlation exists among different phone recognizers. Therefore, not all the phone recognizers are useful for fusion. It is necessary to study how each phone recognizer contributes to the language recognition task.

This paper will take an investigation into the problems stated above. First of all, multiple acoustic diversified phone recognizers are developed to construct the front-end for the PPRVSM system. To get better acoustic diversification, we use different acoustic features and model training methods to develop different phone recognizers. For the acoustic features, the Mel-frequency cepstral coefficients (MFCC) and perceptual linear prediction (PLP) features are both used. In addition, we propose a new time-frequency cepstrum (TFC) feature to extract complementary acoustic information for phone recognition. The TFC feature is obtained by performing a temporal discrete cosine transform (DCT) on the cepstrum matrix and selecting the transformed elements in a specific area with large variances. Recently, feature minimum phone error (fMPE) training has been shown to outperform the conventional maximum-likelihood (ML) training approach [9]. We will examine the use of fMPE training as an alternative to train different acoustic models to capture the acoustic variation within a phone. As the fMPE is performed by transforming the acoustic features, then it can also be treated as a way to get different acoustic features. As mentioned above, it is difficult to combine phone lattices as the acoustic scores in lattices from multiple phone recognizers may not be compatible. This is the case when, for example, combining phone lattices with acoustic scores for ML- and fMPE-trained systems. We propose to fuse the phonotactic features which are extracted from the phone lattices for the acoustic diversified phone recognizers. The fusion of phonotactic features is performed using a simple linear weighting method. The weighting coefficients should be optimized to obtain effective fusion of different features. In this paper, a theoretical inference is made to build a mathematical relationship between the feature weighting coefficients and the score fusion coefficients. And the logistic regression optimized weighting (LROW) method is introduced to optimize the feature weighting coefficients. Since not all the acoustic diversified phone recognizers are effective when fused, we also extend the work by formulating the quantitative measures to select phone recognizers for fusion.

This paper is organized as follows. In Sect. 2, we give a brief description of the PPRVSM system. The details of developing acoustic diversified phone recognizers is described in Sect. 3. Section 4 discusses the fusion technique and selection criterion for effective integration of the acoustic diversified phone recognizers. The experimental results on the NIST 2007 LRE evaluation database are given in Sect. 5, followed by a conclusion in Sect. 6.

2. The PPRVSM System

The phonotactic system employed in this paper for language recognition is the prevailing PPRVSM system, which is depicted in Fig. 1. In the PPRVSM system, a phone lattice is adopted due to its superiority over 1-best phone sequence [7]. According to Fig. 1, the PPRVSM system comprises two main components: the phonotactic feature extraction and the vector space modeling (VSM).

2.1 Phonotactic Feature Extraction

In phonotactic systems, the front-end employs several phone recognizers to convert the speech into phone lattices, which are then used as input to the back-end to perform phonotactic analysis to classify languages. 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. Given the lattice ℓ, the expected counts are calculated over all possible hypotheses in the lattice as follows [7]:

$$\begin{aligned} c(s_i \ldots s_{i+N-1}|\ell) &= \sum_{S\in\ell} p(S|\ell)c(s_i \ldots s_{i+N-1}|S) \\ &= \sum_{s_i \ldots s_{i+N-1}\in\ell} [\alpha(s_i)\beta(s_{i+N-1}) \prod_{j=i}^{i+N-1} \xi(s_j)] \end{aligned}$$

where, $p(S|\ell)$ is the probability of the sequence S in the lattice ℓ, $\alpha(s_i)$ is the forward probability of the starting node in the N-gram $s_i \ldots s_{i+N-1}$, $\beta(s_{i+N-1})$ is the backward probability of the ending node, $\xi(s_j)$ is the posterior probability

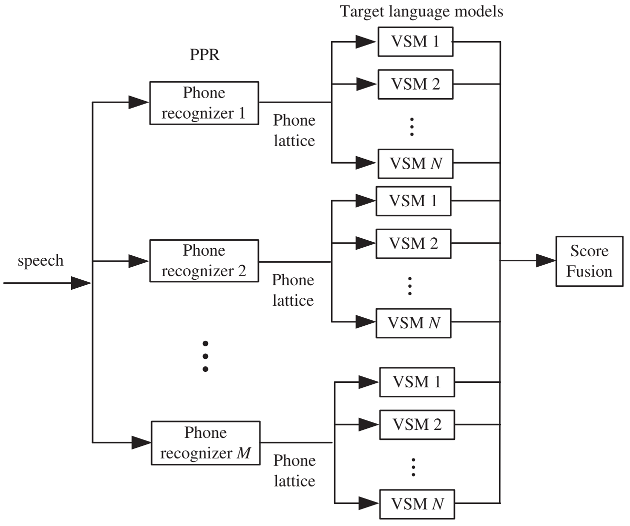

Fig. 1 The PPRVSM system.

of the edge s_j. The probability of the N-gram $s_i \ldots s_{i+N-1}$ in the lattice is then computed as follows:

$$p(s_i \ldots s_{i+N-1}|\ell) = \frac{c(s_i \ldots s_{i+N-1}|\ell)}{\sum_i c(s_i \ldots s_{i+N-1}|\ell)}$$

The probabilities of phonetic N-grams are then concatenated to form a phonotactic feature vector for a given utterance.

2.2 VSM

In VSM, each spoken utterance is represented by a super-vector and then modeled using an SVM [6]. The core of the SVM classifier is the sequence kernel construction, which defines the similarity between two utterances. We employ the term frequency log-likelihood ratio (TFLLR) kernel that has been proven to be effective for phonetic speaker recognition [10]. Let $\vec{X} = \{p(d_1|\ell), \ldots, p(d_F|\ell)\}$ denotes the phonotactic feature for the lattice ℓ, then the kernel between two phonotactic features $\vec{X}_1$ and $\vec{X}_2$ is

$$\begin{aligned} K(\vec{X}_1, \vec{X}_2) &= \sum_{i=1}^{F} p_n(d_i|\ell_1) * p_n(d_i|\ell_2) \\ &= \sum_{i=1}^{F} \frac{p(d_i|\ell_1)}{\sqrt{p(d_i|all)}} * \frac{p(d_i|\ell_2)}{\sqrt{p(d_i|all)}} \end{aligned} \tag{1}$$

where, $d_i = s_i \ldots s_{i+n-1}$ $(n \leq N)$ and $F = f + f^2 + \cdots + f^N$ (f is the size of the phone inventory for a single phone recognizer). The denominator $p(d_i|all)$ is the probability of d_i in all the phone lattices used for training, which is chosen to make sure that N-grams with large probabilities will not dominate the similarity in the kernel. The inner product form in Eq. (1) indicates that a high degree of similarity will exist between the two lattices if the same N-grams are present in them two, and vice versa.

Given the scaled super-vector $\vec{X}$ and the kernel function $K(\vec{X}, \vec{X}_l)$, the SVM scoring can be implemented as follows:

$$f(\vec{X}) = \sum\nolimits_l \alpha_l K(\vec{X}, \vec{X}_l) + d \tag{2}$$

A decision is based on the output of the SVM in Eq. (2) compared to a threshold. The $\vec{X}_l$ are support vectors trained using the Mercer condition. The training is carried out with a one-versus-rest strategy. We view the samples in the target language as the positive set and the remainder as the negative one. The training is carried out between them two.

Generally, high-order N-grams have higher discriminative ability for language recognition compared with low-order ones [6]. But it is problematic that the number of N-grams grows exponentially as the order N increases. Then a high dimensional phonotactic feature vector will be produced if we use high-order N-grams, which is a challenge for SVM training. However, experiments have demonstrated that not all the N-grams are necessary [11]. In this paper, we use a method similar to the discriminative keyword selection proposed in [11] to pick out the most discriminative N-grams for language recognition. The approach comprises two stages: selection and construction. The selection is designed to pick out the most discriminative low-order N-grams using the max-relevance criteria based on mutual information, the construction is the process of building high-order N-grams based on the selected low-order ones. Details can be found in our previous work [12].

3. Development of the PPR Front-End Based on Acoustic Diversified Phone Recognizers

3.1 Acoustic Diversified Phone Recognizers

There are two main problems for phonotactic systems using only one phone recognizer. First, the recognizer is language specific. Then the phone set may not cover the sound units of another language. It makes sense to employ multiple phone recognizers of different languages to broaden the phone coverage. The other problem is that the phone recognition results are error-prone, which will inevitably degrade the performance of the back-end phonotactic language models. This can be solved by combining the results from multiple acoustic diversified phone recognizers which is similar to [8]. Using phone recognizers of different languages often comes at the cost of collecting multiple sets of annotated training samples. The introduction of acoustic diversified phone recognizers is attractive as it maximizes the use of limited transcribed training data. But it is a challenge to develop phone recognizers that will extract complementary information for language classification.

The construction of acoustic diversified phone recognizers can be implemented by training different acoustic models for phone recognizers using the same training data and phone set. Generally, the phone lattices obtained by acoustic diversification are homogeneous as the same phone set is used. It is necessary to train different acoustic models that will generate phone lattices containing error patterns complementary to each other. Simple methods such as simply changing the number of model parameters are infeasible, they yield only slight changes in phone lattices. Therefore, different acoustic features and model training paradigms are favorable to construct complementary phone recognizers on the same speech corpus. In this paper, the following techniques are employed to build the acoustic diversified PPR front-end:

(1) Acoustic features. Usually, MFCC and PLP are used for phone recognition in the PPRVSM system. In our previous work, we have presented a time-frequency cepstrum (TFC) feature for the GMM-based acoustic system [13], which utilizes a temporal discrete cosine transform (DCT) on the cepstrum matrix and outperforms the widely used shifted delta cesptrum (SDC) feature. We will extend this work for phone recognition in the PPRVSM system.

(2) Model training methods. Recently, discriminative training has become an attractive technique as it outperforms the conventional ML training approach in speech recognition, such as Minimum Phone Error (MPE) [14] and fMPE [9]. In application, all techniques are performed on acoustic models except fMPE, which applies to the acoustic

features. We will adopt the fMPE technique for discriminative training of phone models.

Six phone recognizers are developed by using the techniques stated above: (1) MA-MFCC-ML; (2) MA-MFCC-fMPE; (3) MA-PLP-ML; (4) MA-PLP-fMPE; (5) MA-TFC-ML; (6) MA-TFC-fMPE. They are all trained on Mandarin speech data and modeled using Gaussian Mixture Model/Hidden Markov Model (GMM/HMM). But different acoustic features (MFCC, PLP, TFC) and model training methods (ML, fMPE) are adopted. For example, MA-MFCC-ML means the GMM/HMM structured Mandarin phone recognizer uses MFCC as the acoustic feature and ML as the training method. The fundamentals of the TFC feature and fMPE training will be introduced next.

3.2 TFC Feature Extraction

In our previous work, we have proposed the TFC feature in the GMM-based acoustic system for language recognition [13]. The extraction is performed as follows: several successive frames of basic feature vectors within a context width are first extracted to form a cepstrum matrix. A temporal DCT is then performed on the cepstrum matrix to remove the correlation in the temporal direction. Finally, the elements in the upper-left triangular area are selected in a zigzag scan order.

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 due to small scale acoustic events. Also, the elements can be selected with a greater variability for the TFC feature.

In the GMM-based acoustic systems, the context width is about 20 frames. The normalized variances of the cepstrum matrix after a horizontal DCT is nearly triangular, thus we can perform a zigzag scan to select elements in this area to form the TFC feature. But for phone recognition, the optimal configuration will not be the same as that for the GMM system. To show this, the variance of each element in the cepstrum matrix (using successive 9 frames of 20-dimensional MFCC basic feature vector) after a temporal DCT was computed on the data corpus used for training phone models. The normalized variances (normalized by the maximum elements) are plotted in Fig. 2. We can see that there are other effective configurations besides the triangle adopted in TFC, such as a rectangle.

3.3 fMPE

The fMPE discriminative training method employs the same objective function as MPE, which can be written as:

$$F(\lambda) = \sum_{r=1}^{R} \sum\nolimits_{s} P_{\lambda}^{k}(s|O_r) A(s, s_r)$$

where $P_{\lambda}^{k}(s|O_r)$ is the scaled posterior probability of hypothesis s given the r'th observation O_r, k is the acoustic scaling factor and λ is the current model vector. The $A(s, s_r)$ is the raw phone accuracy between hypothesis s and reference transcription s_r. This criterion is an average of the transcription accuracies of all possible sentences s.

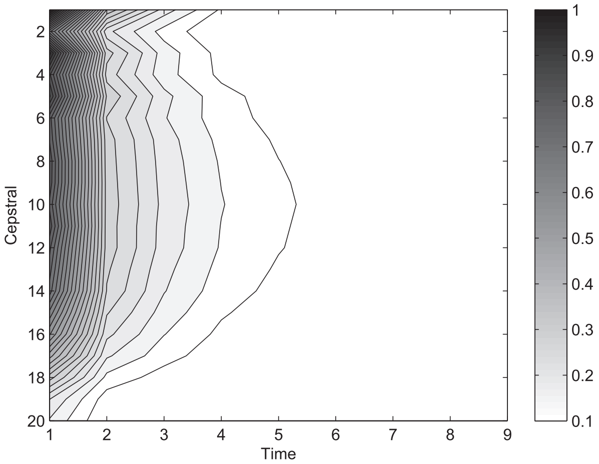

Fig. 2 The normalized variances of each element in the cepstrum matrix after a horizontal DCT.

The implementation of fMPE is carried out by transforming the acoustic feature with a kernel-like method, where offsets to the features are obtained by training a projection from a high-dimension feature space based on posteriors of Gaussians [9]. Let x_t be the original features and y_t be the transformed features. The transformation is:

$$y_t = x_t + Mh_t$$

where h_t is the expanded high dimensional feature derived from posteriors of Gaussians. M is the transform matrix that needs to be estimated to optimize the MPE objective function. The detailed calculation method of h_t and M can be found in [9].

As with normal fMPE training in speech recognition, we need to generate lattices by decoding the training data with a weak language model [15], which are used to produce the MPE statistics. For experiments, we use the more robust and effective offset features described in [16] to obtain the high dimension vector h_t. In this study, 1000 Gaussians are used to calculate the offset features with context width 5. Typically, we run 3-4 iterations of fMPE optimization.

4. Fusion and Selection of the Acoustic Diversified Phone Recognizers

Generally, the PPRVSM system comprises multiple parallel subsystems, where each subsystem employs a phone recognizer to extract phonotactic attributes from the speech to characterize a language. Each phone recognizer extracts complementary information from the speech so that improvements can be attained when fusing the subsystems. Usually, score fusion is adopted because it doesn't require detailed knowledge of the structure of the subsystems for

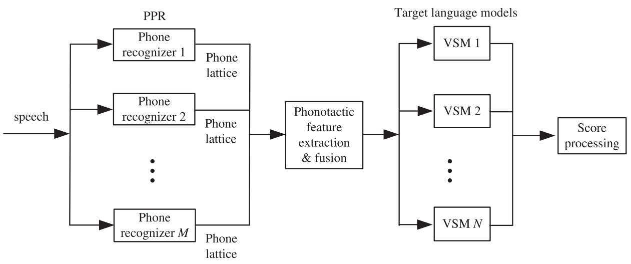

Fig. 3 The PPRVSM system based on feature fusion.

fusion and is easy to implement. Actually, the fusion can also be performed at the feature or model level. From an information theoretic perspective, the amount of information will decrease with each operation in a system. Then Fusion at the feature or model level can theoretically retain more information compared to fusion at the score level. In this paper, we will adopt feature fusion for the acoustic diversified phone recognizers. Figure 3 shows the structure of the PPRVSM system using the feature fusion method.

4.1 Fusion of Phonotactic Features

In the PPRVSM system, the classical method of feature fusion is to group several sets of phonotactic feature vectors resulting from the PPRs into a large composite supervector [6]. However, this increases the dimension so more training data is required to ensure robust estimation of model parameters. In this paper, we propose a novel strategy for phonotactic feature fusion. The idea is to map multiple feature vectors of different phone recognizers into one through a predefined transformation.

Suppose $\vec{X}_1, \ldots, \vec{X}_m$ are the different phonotactic features for an input utterance x. The concatenated feature vector $\vec{X}$ can be represented as $\vec{X} = \left(\vec{X}_1^T, \ldots, \vec{X}_m^T\right)^T$. Denoting each vector $\vec{X}_m$ as d_m-dimensional, the concatenated feature $\vec{X}$ is $(d_1 + \cdots + d_m)$-dimensional. Here, we intend to transform these vectors into one, rather than concatenate them. The mapping is defined by

$$\vec{X} = f\left(\vec{X}_1, \ldots, \vec{X}_m\right)$$

The transformation function can be linear or non-linear, just as with score fusion. There are already a variety of score fusion techniques, including SVM, linear discriminate analysis followed by Gaussian mixture models (LDA+GMM) and linear fusion [17]. One recent experiment [18] has shown that the linear fusion method is not only simple to achieve but also yields good results in practice. Although other methods have been able to give a more detailed portrait of the scores, they come at the expense of the robustness of system in complex conditions. Therefore, we will adopt the linear fusion method to fuse the phonotactic features from different phone recognizers.

Let ℓ_m denote the phone lattice decoded using the m-th phone recognizer. Then the phonotactic feature vector of x estimated from the m-th phone recognizer can be defined as $\vec{X}_m = \{p(d_1|\ell_m), \ldots, p(d_F|\ell_m)\}$. The goal of linear fusion is to find a set of weighting coefficients $\{w_1, w_2, \ldots, w_M\}$ such that the fusion of M phone recognizers

$$\vec{X} = \sum_{i=1}^{M} w_i \vec{X}_i \tag{3}$$

has the maximum discriminative ability. For simplicity, the average weighting method can be adopted, where the weight coefficients are equal. However, this is not the optimal method when the phone recognizers have different contributions to fusion. A weighting optimization method is favorable.

Recently, researchers have proposed an optimization criterion for linear score fusion based on the cost of log-likelihood rate (CLLR) and implemented using logistic regression [19]. Because of its superior performance, this method has gradually become the primary method of linear fusion. It can be formulated as follows. For the trial x, the scores of K target languages produced by the m-th subsystem can be denoted as $\vec{s}_m(x) = [s_{m1}(x), \ldots, s_{mK}(x)]$. A transformation function is then defined in the following:

$$\vec{s}(x) = \sum_{i=1}^{M} w_i' \vec{s}_i(x) \tag{4}$$

where w_m' is a scalar, $\{w_1', w_2', \ldots, w_M'\}$ constitutes a set of transformation parameters which are estimated using a conjugate gradient descent algorithm to minimize a linear regression function [20], which we use here. The fusion factor w_m' implies the role of the subsystem in classification. The larger w_m' is, the more important the subsystem is and the more discriminative the feature is.

Although w_m' reflects the role of the phonotactic features indirectly, we can't use it to replace w_m in Eq. (3). There exists a relationship between the score fusion weights $\{w_1', w_2', \ldots, w_M'\}$ and the phonotactic feature fusion weights $\{w_1, w_2, \ldots, w_M\}$. Expanding Eq. (2):

$$\begin{aligned} s_k(x) &= \sum_t \alpha_{kt} K(\vec{X}, \vec{X}_{kt}) + d \\ &= \sum_t \alpha_{kt} \left[\left(\sum_{i=1}^{M} w_i b(\vec{X}_i)\right) * \left(\sum_{i=1}^{M} w_i b(\vec{X}_{kti})\right)\right] + d \end{aligned} \tag{5}$$

where the α_{kt} and X_{kt} are the parameters of the k-th language model. We assume that the phonotactic features extracted using different phone recognizers are complementary to each other, then the inner products of them are approximately zero. Also, the bias d is constrained to be zero during SVM training. Then we get the following expression from Eq. (5):

$$\begin{aligned} s_k(x) &= \sum_t \alpha_{kt} \left[\sum_{i=1}^{M} w_i^2 b(\vec{X}_i) b(\vec{X}_{kti})\right] \\ &= \sum_{i=1}^{M} w_i^2 \sum_t \alpha_{kt} b(\vec{X}_i) b(\vec{X}_{kti}) = \sum_{i=1}^{M} w_i^2 s_{ik}(x) \end{aligned} \tag{6}$$

Combining Eqs. (4) and (6) leads to the approximation:

$$w_i \approx \sqrt{w_i^{'}} \tag{7}$$

To implement the logistic regression optimized weighting method, we first estimate the score transformation parameters $\{w_1^{'}, w_2^{'}, \ldots, w_M^{'}\}$ using the logistic regression in the score domain on the development set. Then we map the $\{w_1^{'}, w_2^{'}, \ldots, w_M^{'}\}$ back to the feature domain using Eq. (7). This kind of score fusion aims at optimizing the mixture-of-experts system by calibrating the parameter set. For score fusion using logistic regression, we can see from Eq. (4) that the fusion factor $w_m^{'}$ should be approximately 0 if the m-th subsystem is a nuisance factor and greater than 0 otherwise. Thus useless subsystems are filtered out automatically. However, empirical results show that limited training data generally results in non-zero fusion factors for all classifiers, increasing system complexity and potentially hurting results. This can be addressed through the use of a pre-fusion subsystem selection process [21]. Next, we will investigate into selecting acoustic diversified phone recognizers.

4.2 Selection Strategy for the Acoustic Diversified Phone Recognizers

While more phone recognizers might not harm the system, we may not wish to use all the phone recognizers in the front-end. First, the phone sets of these acoustic diversified phone recognizers are the same. The phone lattices are often highly overlapped even if different acoustic features or model training techniques are used, resulting in highly correlated language recognition scores. Second, more phone recognizers mean higher computational cost. In view of this, our strategy is to select phone recognizers that have high discriminative ability for target languages and have little redundancy.

Of course, we can select m complementary phone recognizers from M phone recognizers for fusion using a combinatorial method, with a total number of combinations equal to C_M^m. However, it is laborious to examine through all combinations to find the global optimum solution. In this paper, we adopt a criterion to measure the merit of different acoustic diversified phone recognizers similarly to [22], which is based on the discriminative ability and distinctiveness of a phone recognizer. We will discuss the criterion.

1. Discriminative ability of phone recognizers: In the PPRVSM framework, the phone recognizers are used to convert the speech into a set of phone lattices, from which the phonetic N-gram statistics are estimated to model the languages. Then the entropy of the N-gram statistics can be used to evaluate the discriminative ability of the phone lattices in language recognition [23], [24]. The conditional entropy of N-gram statistics relative to K target languages can be calculated as

$$H(\vec{X}_m|L) = -\sum_{k=1}^{K}\sum_{i=1}^{F_m} p(d_i, l_k)\log p(d_i|l_k) \tag{8}$$

where L denotes a set of target languages $L = \{l_1, l_2, \ldots, l_K\}$ and K is the number of target languages. d_i are the phonetic N-grams and F_m is the number of different N-grams. From Eq. (8), we can see that lower conditional entropy means less uncertainty about the language identity. Phone recognizers with lower conditional entropy will have higher discriminative ability for language recognition.

2. Distinctiveness of phone recognizers: The pair-wise hamming distance of phone sequences based on binary code is employed to measure the distinctiveness of different tokenizers derived from the same phone recognizers in [22]. In this paper, we use phone lattice instead of sequence. Then we adopt the Euclidean distance between the "seen" phonetic N-gram statistics to evaluate the distinctiveness of different phone recognizers, which is defined below:

$$D_m = \frac{1}{(M-1) * F_m}\sum_{j\neq m}^{M}\sum_{i=1}^{F_m}\sqrt{(c_m(d_i) - c_j(d_i))^2}$$

where $c(d_i)$ are the expected counts of the N-gram statistics which has been defined in Sect. 2.1 and M is the number of phone recognizers. The greater D_m is, the more distinctive the phone recognizer is from others.

A good phone recognizer should have high distinctiveness value and low conditional entropy. Therefore, the final evaluation of different phone recognizers can be formulated as follows:

$$E_m = \frac{D_m}{H(\vec{X}_m|L)}$$

We rank the acoustic diversified phone recognizers according to the evaluation results. The phone recognizers with the largest values are selected to form the PPR front-end in PPRVSM language recognition system.

5. Experiments

5.1 Experimental Setup

The experiments are performed on the NIST 2007 LRE evaluation database under both closed-set and open-set test condition. For each trial, if the set of non-target languages will be the set of LRE 2007 target languages, minus the target language, this is the "closed-set" test condition. If the set of non-target languages also includes other "unknown" languages whose identities will not be disclosed, this is the "open-set" test condition. The task of LRE 2007 is to recognize 14 languages: Arabic, Bengali, Chinese, English, Farsi, German, Hindustani, Japanese, Korean, Russian, Spanish, Tamil, Thai and Vietnamese. There are 7530 utterances in total, spanning the 3, 10 and 30 second conditions [25]. As the evaluation corpus include some out-of-set languages, only 6474 utterances are used for the closed-set test condition and all 7530 utterances are used for the open-set test condition. The training data comes from different sources including Callfriend, the development and evaluation data provided by NIST in the previous LRE. The Callfriend corpus and the development data of previous LRE are employed to select phonotactic features and train VSM. The evaluation data from previous LRE are used for score fusion and estimating the weighting coefficients of LROW. Indeed, it will be better to use different data for creating the VSM and selecting phonotactic feature. But there are not enough data for some languages such as Bengali. So we use all data for both VSM modeling and feature selection. As the speech is relatively long, we use voice activity detection to segment each speech utterance into segments, which are about 30 seconds in length.

In the PPRVSM language recognition, several phone recognizers are used in parallel to decode the speech into phone lattices for analysis. The Mandarin phone recognizers employed in our experiments are developed using the GMM/HMM architecture and trained on about 30 hours of conversational telephone data. There are 64 phone models for the phone recognizer, each of which is a tied-state left-to-right context-dependent GMM/HMM with 32 Gaussians per state. For acoustic feature extraction, 12 MFCC coefficients 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 (including energy). To remove channel variability, cepstral mean subtraction and variance normalization are both applied. In addition to the MFCC, the PLP feature is also extracted to provide additional acoustic diversification. The parameters and configuration for PLP extraction are similar to the MFCC feature except that c_0 is used instead of energy. For TFC feature extraction, the optimal configuration is determined empirically to be complementary to MFCC and PLP features, as described in the next section. For phonotactic feature selection, the top 20% of the low-order N-grams are selected based on mutual information [12].

5.2 Experimental Results

We demonstrate the effectiveness of our approaches under 3, 10 and 30 second conditions. Both pooled equal error rate (EER) and detection cost function (DCF, Cavg*100) are used to summarize the results, which are obtained by pooling the scores of all target (or non-target) languages together.

The first experiment is to find the optimal configuration of the reserved area for TFC feature extraction. The context width is fixed to 9. In this work, we adopt a rectangular shape 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}). Besides the TFC, N-dimension static MFCC parameters are also appended. The language recognition results are given in Tables 1 and 2. When the EER and the Cavg*100 are not consistent, we consider the Cavg*100 as the primary metric. From these tables, we can see that we will get the best performance using 52 dimensions (39 + 13 static parameters). We adopt the 52 dimension TFC to serve as an alternative feature for phone recognition as the dimension and the derivation of the elements are complementary to both MFCC and PLP.

Our second experiment is to show the performance of the acoustic diversified phone recognizers. Tables 3 and 4 summarize the language recognition results using different phone recognizers in the front-end. The abbreviation for each phone recognizer is defined in Sect. 3.1. From Tables 3 and 4, we can see that these phone recognizers are comparable in performance. Since they use different acoustic features and model training paradigms, a complementary effect is expected when fusion is performed.

The third experiment is implemented to compare different methods of fusing multiple acoustic diversified phone

Table 1 Comparison of TFC with different rectangular area under the closed-set test condition.

EER / Cavg*100	30 second	10 second	3 second
TFC $13 \times 2 + 13$	3.09 / 2.83	9.67 / 9.38	21.14 / 20.53
TFC $13 \times 3 + 13$	2.60 / 2.43	8.61 / 8.22	20.10 / 19.73
TFC $13 \times 4 + 13$	2.54 / 2.45	8.58 / 8.34	20.74 / 20.10

Table 2 Comparison of TFC with different rectangular area under the open-set test condition.

EER / Cavg*100	30 second	10 second	3 second
TFC $13 \times 2 + 13$	4.33 / 5.23	10.85 / 11.87	21.60 / 21.86
TFC $13 \times 3 + 13$	3.67 / 4.72	9.59 / 10.89	20.47 / 21.25
TFC $13 \times 4 + 13$	3.64 / 4.62	9.74 / 10.76	21.35 / 21.40

Table 3 Performance of different acoustic diversified phone recognizers under the closed-set test condition.

EER / Cavg*100	30 second	10 second	3 second
MA-MFCC-ML	2.93 / 2.63	9.60 / 8.88	21.34 / 21.01
MA-MFCC-fMPE	2.65 / 2.38	8.90 / 8.41	20.25 / 20.24
MA-PLP-ML	2.69 / 2.64	8.76 / 8.48	20.63 / 20.18
MA-PLP-fMPE	2.42 / 2.35	8.64 / 8.24	19.82 / 19.28
MA-TFC-ML	2.60 / 2.43	8.61 / 8.22	20.10 / 19.73
MA-TFC-fMPE	2.45 / 2.35	8.24 / 8.07	20.56 / 19.77

Table 4 Performance of different acoustic diversified phone recognizers under the open-set test condition.

EER / Cavg*100	30 second	10 second	3 second
MA-MFCC-ML	3.84 / 4.80	10.37 / 11.34	21.89 / 22.42
MA-MFCC-fMPE	3.71 / 4.53	9.93 / 11.08	20.87 / 21.38
MA-PLP-ML	3.74 / 4.77	9.82 / 11.08	21.81 / 20.18
MA-PLP-fMPE	3.41 / 4.45	9.71 / 10.68	20.65 / 20.72
MA-TFC-ML	3.67 / 4.72	9.59 / 10.89	20.47 / 21.25
MA-TFC-fMPE	3.52 / 4.36	9.24 / 10.40	21.17 / 21.30

Table 5 Comparison of different methods of fusing multiple acoustic diversified phone recognizers under the closed-set test condition.

EER / Cavg*100	30 second	10 second	3 second
Score fusion	2.09 / 2.04	7.18 / 6.89	19.28 / 18.83
Feature fusion (AW)	1.93 / 1.86	6.92 / 6.62	17.67 / 17.40
Feature fusion (LROW)	1.62 / 1.53	6.89 / 6.52	17.69 / 17.28

Table 6 Comparison of different methods of fusing multiple acoustic diversified phone recognizers under the open-set test condition.

EER / Cavg*100	30 second	10 second	3 second
Score fusion	3.01 / 4.11	8.15 / 9.67	19.88 / 20.49
Feature fusion (AW)	3.03 / 3.80	7.98 / 9.29	18.28 / 19.02
Feature fusion (LROW)	2.69 / 3.52	7.99 / 8.95	18.64 / 18.93

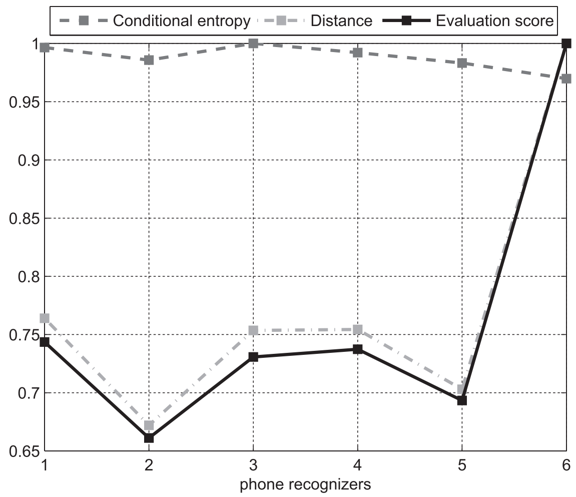

Fig. 4 Different evaluation criteria for the acoustic diversified phone recognizers.

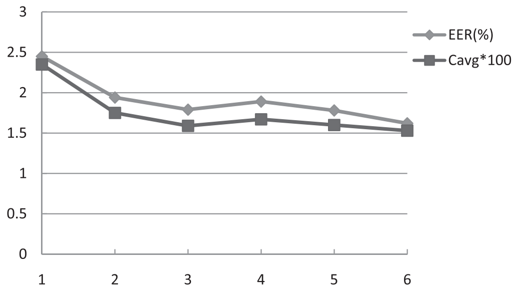

Fig. 5 Performance as a function of the number of phone recognizers (30 second, closed-set).

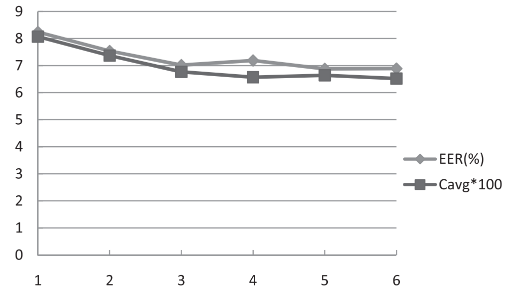

Fig. 6 Performance as a function of the number of phone recognizers (10 second, closed-set).

recognizers. The results are summarized in Tables 5 and 6. The AW and LROW in these tables mean average weighting and logistic regression optimized weighting respectively. The LDA+GMM is used for score fusion. From Tables 5 and 6, we can see that improvements can be attained when fusing different phone recognizers. Moreover, fusing features is more effective than fusing scores, which is consistent with our assumption proposed in Sect. 4: more information can be utilized if we fuse at the feature or model level rather than at the score level. Since the LROW method focuses on minimizing the CLLR, better weighting coefficients can be obtained compared with AW. This can also be observed from Tables 5 and 6. However, the performance improvements are more significant for the 30 second test set compared with the 10 second and 3 second test set. That is because the LROW method is performed on the 30 second development set. More improvements can be attained for the 10 second and 3 second test set if we optimize the coefficients on the 10 second and 3 second development set respectively. In comparison with the score fusion method, the LROW based feature fusion method achieves a relative decrease of 22.49%, 4.04% and 8.25% in EER for the 30 second, 10 second and 3 second closed-set test conditions separately.

Next, we carried out experiments to examine the optimal number of acoustic diversified phone recognizers for a PPR front-end. As discussed in Sect. 4.2, we calculated the conditional entropy and the distance of each phone recognizer for evaluation. The normalized values (normalized by the maximum elements) are plotted in Fig. 4, where the two dotted lines show the discriminative ability and distinctiveness of each phone recognizer and the solid curve is used for overall evaluation. Although some phone recognizers have high discriminative abilities, their evaluation scores are low due to low distinctiveness. Our strategy was to select phone recognizers with the largest evaluation values to build the PPRVSM front-end, which will maintain the best discriminative ability with reduced redundancy. We first ranked and short-listed the phone recognizers by their evaluation values to form the PPR front-end. Figures 5–10 show the performance achieved by using different number of phone recognizers on the 30 second, 10 second and 3 second test set. The x-axis of each figure indicates the number of phone recognizers used to form a fused system using the LROW based feature fusion method. One can observe that, as the number of phone recognizers increases, the language recognition performance is improved and saturates after three phone recognizers is employed. Our proposed PPRVSM system will be built by fusing the selected three acoustic diversified phone recognizers at the feature level. The EERs of the system achieves 1.79%, 7.02% and 18.02% for the 30 second, 10 second and 3 second closed-set test conditions respectively, which are comparable with the results in the last row

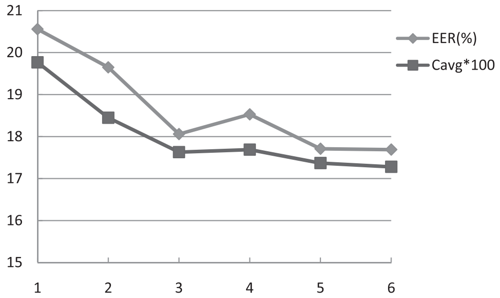

Fig. 7 Performance as a function of the number of phone recognizers (3 second, closed-set).

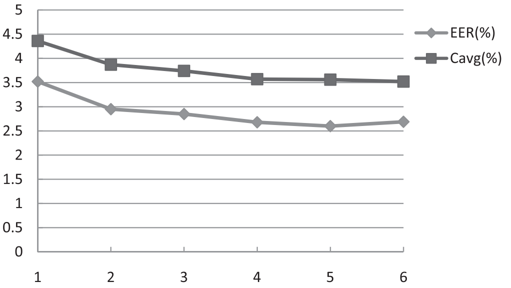

Fig. 8 Performance as a function of the number of phone recognizers (30 second, open-set).

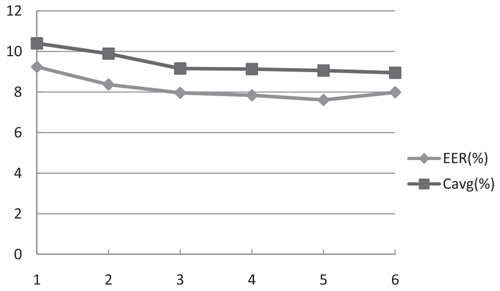

Fig. 9 Performance as a function of the number of phone recognizers (10 second, open-set).

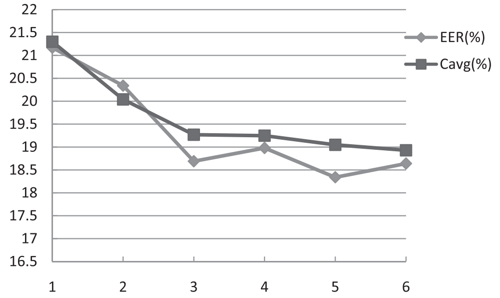

Fig. 10 Performance as a function of the number of phone recognizers (3 second, open-set).

of Table 5.

Finally, we compare the proposed PPRVSM system with a traditional one. In the traditional PPRVSM system, three phone recognizers with phonetic diversification are employed, which are developed by the Faculty of Information Technology of the Brno University of Technology [26]. They are trained on different languages: Czech, Hungarian and Russian. For each language, about 10 hours of speech are used in training. They all adopt the Artificial Neural Network/Hidden Markov Model (ANN/HMM) structure. The number of each phone set is: 42 for the Czech, 58 for the Hungarian and 49 for the Russian. In Tables 7 and 8, the proposed PPRVSM (a) is composed of three selected acoustic diversified phone recognizers stated above, which are fused at the feature level using LROW. The traditional PPRVSM (b) is built by fusing the Czech, Hungarian and Russian phone recognizers at the score level. The results in Tables 7 and 8 show that our proposed PPRVSM (a) is competitive with the traditional one (b) for the three test conditions. The results are significantly better for the 30 second test set. This fact illustrates that the LROW method performs better using a matched length of speech for optimization. We also conduct a score fusion experiment using the two systems (a+b). The LDA+GMM method is adopted for score fusion. The further improvement indicates that the two systems with different diversifications are complementary to each other. This is likely because the Czech, Hungarian and Russian phone recognizers use the ANN/HMM architecture while the Mandarin phone recognizer adopts a different GMM/HMM architecture. The EERs of the fused system achieve 1.24%, 4.98% and 14.96% for the 30 second, 10 second and 3 second closed-set test conditions respectively.

Table 7 Comparison and fusion of different PPRVSM systems under the closed-set test condition.

EER / Cavg*100	30 second	10 second	3 second
The proposed PPRVSM (a)	1.79 / 1.59	7.02 / 6.77	18.06 / 17.63
The traditional PPRVSM (b)	2.59 / 2.29	7.41 / 7.63	19.07 / 17.99
a+b	1.24 / 1.13	4.98 / 4.92	14.96 / 14.75

Table 8 Comparison and fusion of different PPRVSM systems under the open-set test condition.

EER / Cavg*100	30 second	10 second	3 second
The proposed PPRVSM (a)	2.85 / 3.74	7.96 / 9.16	18.69 / 19.27
The traditional PPRVSM (b)	3.67 / 4.73	8.50 / 9.79	19.70 / 19.56
a+b	2.17 / 2.95	6.07 / 7.39	15.68 / 16.59

6. Conclusions

We investigate a strategy to build a PPRVSM system for language recognition based on parallel phone recognizers that are acoustically diversified and fused at the feature level. A variety of acoustic features and model training methods are employed to develop the diversified recognizers. Features include MFCC and PLP along with a novel time-frequency feature are studied to provide complementary acoustic emphasis. We propose a new technique based on LROW for effective fusion of different phonotactic features. We also introduce a method to study how each diversified phone recognizer contributes to the language recognition task, which helps determine the number of phone recognizers to include.

There are two main advantages of the proposed system: (1) additional phone recognizers can be built without the need for additional annotated speech samples; (2) fusing at the feature level retains more information than fusion at the score level. Experimental results on the NIST 2007 LRE evaluation set show that the proposed system is comparable and complementary to an established PPRVSM system for phonotactic language recognition. When we fuse the two systems at the score level for further improvements, the EERs achieve 1.24%, 4.98% and 14.96% for the 30 s, 10 s and 3 s closed-set test conditions respectively.

Acknowledgments

The authors would like to thank Dr. Michael T. Johnson for fruitful discussions and assistance with English writing. This work is supported by National Natural Science Foundation of China and Microsoft Research Asia (60776800), National Natural Science Foundation of China and Research Grants Council (60931160443), National Natural Science Foundation of China (61005019), National High Technology Research and Development Program of China (863 Program) (2008AA02Z414, 2008AA040201).

References

[1] Y.K. Muthusamy, E. Barnard, and R.A. Cole, "Reviewing automatic language identification," IEEE Signal Process. Mag., vol.11, no.4, pp.33–41, 1994.

[2] M.A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," IEEE Trans. Speech Audio Process., vol.4, no.1, pp.31–44, 1996.

[3] 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," Proc. ICSLP, pp.33–36, Denver, Colorado, Sept. 2002.

[4] W.M. Campbell, J.P. Campbell, D.A. Reynolds, E. Singer, and P.A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," Comput. Speech Lang., vol.20, no.2-3, pp.210–229, 2006.

[5] J. Navratil, "Recent advances in phonotactic language recognition using binary decision trees," Proc. ICSLP, pp.421–424, Pittsburgh, Sept. 2006.

[6] H. Li, B. Ma, and C-H. Lee, "A vector space modeling approach to spoken language identification," IEEE Trans. Audio, Speech Language Process., vol.15, no.1, pp.271–284, 2007.

[7] J.L. Gauvain, A. Messaoudi, and H. Schwenk, "Language recognition using phone lattices," Proc. ICSLP, pp.1283–1286, Jeju Island, Oct. 2004.

[8] K.C. Sim and H. Li, "On acoustic diversification front-end for spoken language identification," IEEE Trans. Audio, Speech Language Process., vol.16, no.5, pp.1029–1037, 2008.

[9] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "fMPE: Discriminatively trained features for speech recognition," Proc. ICASSP, pp.961–964, Philadelphia, PA, March 2005.

[10] W.M. Campbell, J.P. Campbell, D.A. Reynolds, D.A. Jones, and T.R. Leek, "Phonetic speaker recognition with support vector machines," in Advances in Neural Information Processing System 16, Eds. S. Thrun, L. Saul, and B. Scholkopf, pp.1377–1384, MIT Press, Cambrige, MA, 2004.

[11] F.S. Richardson and W.M. Campbell, "Language recognition with discriminative keyword selection," Proc. ICASSP, pp.4145–4148, Las Vegas, Nevada, 2008.

[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, no.supp1, pp.138–144, 2009.

[13] 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. Audio, Speech Language Process., accepted, 2010.

[14] D. Povey, Discriminative training for large vocabulary speech recognition, Ph.D. dissertation, Cambridge University, Cambridge, UK, 2004.

[15] R. Schluter, B. Muller, F. Wessel, and H. Ney, "Interdependence of language model and discriminative training," Proc. ASRU Workshop, pp.119–122, Keystone, Colorado, Dec. 1999.

[16] D. Povey, "Improvements to fMPE for discriminative training of features," Proc. Interspeech, pp.2977–2980, Lisbon, Sept. 2005.

[17] W.M. Campbell, D.A. Reynolds, and J.P. Campbell, "Fusing discriminative and generative methods for speaker recognition: Experiments on switchboard and NFIITNO field data," Proc. Odyssey04, pp.41–44, Toledo, 2004.

[18] D. Leeuwen and N. Brummer, "Channel-dependent GMM and multi-class logistic regression models for language recognition," Proc. Odyssey06, San Juan, June 2006.

[19] N. Brummer and J. Preez, "Application-independent evaluation of speaker detection," Comput. Speech Lang., vol.20, no.2-3, pp.230–275, 2006.

[20] N. Brummer and D. Leeuwen, "On calibration of language recognition scores," Proc. Odyssey06, San Juan, June 2006.

[21] T. Hou and J. Liu, "Vector angle minimum criteria for classifier selection in speaker verification technology," Chinese Journal of Electronics, vol.19, no.1, pp.81–85, 2010.

[22] R. Tong, B. Ma, H. Li, and E.S. Chng, "A target-oriented phonotactic front-end for spoken language recognition," IEEE Trans. Audio, Speech Language Process., vol.17, no.7, pp.1335–1347, 2009.

[23] J. Navratil, "Spoken language recognition — A step toward multilinguality in speech processing," IEEE Trans. Speech Audio Process., vol.9, no.6, pp.678–685, Aug. 2001.

[24] R. Rosenfeld, Adaptive Statistical Language Modeling: A Maximum Entropy Approach, Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, PA, 1992.

[25] "The 2007 NIST Language Recognition Evaluation Plan," http://www.itl.nist.gov/iad/mig/tests/lre/2007/L-RE07EvalPlan-v8b.pdf, 2007.

[26] P. Matejka, P. Shwarz, J. Cernocky, and P. Chytil, "Phonotactic language identification using high quality phoneme recognition," Proc. Interspeech, pp.2237–2240, Lisbon, Portugal, Sept. 2005.

Yan Deng was born in 1982. She received the B.S. degree in communication engineering from National University of Defense Technology, Changsha, China, in 2005. She is currently pursuing the Ph.D. degree in the Department of Electronic Engineering, Tsinghua University, Beijing, China. Her research focuses upon language recognition and speaker recognition.

Wei-Qiang Zhang 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 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 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 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.