State-Dependent Phoneme-Based Model Merging for Dialectal Chinese Speech Recognition

Linquan Liu, Thomas Fang Zheng, and Wenhu Wu

Center for Speech Technology, Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing, 100084
liulq@cst.cs.tsinghua.edu.cn, fzheng@tsinghua.edu.cn, wuwh@tsinghua.edu.cn

Abstract. Aiming at building a dialectal Chinese speech recognizer from a standard Chinese speech recognizer with a small amount of dialectal Chinese speech, a novel, simple but effective acoustic modeling method, named *state-dependent phoneme-based model merging* (SDPBMM) method, is proposed and evaluated, where a tied-state of standard triphone(s) will be merged with a state of the dialectal monophone that is identical with the central phoneme in the triphone(s). It can be seen that the proposed method has a good performance however it will introduce a Gaussian mixtures expansion problem. To deal with it, an acoustic model distance measure, named *pseudo-divergence based distance measure*, is proposed based on the difference measurement of Gaussian mixture models and then implemented to downsize the model size almost without causing any performance degradation for dialectal speech. With a small amount of only 40-minute Shanghai-dialectal Chinese speech, the proposed SDPBMM achieves a significant absolute syllable error rate (SER) reduction of 5.9% for dialectal Chinese and almost no performance degradation for standard Chinese. In combination with a certain existing adaptation method, another absolute SER reduction of 1.9% can be further achieved.

Keywords: Speech recognition, dialectal Chinese speech recognition, state-dependent phoneme-based model merging, acoustic modeling, acoustic model distance measure.

1 Introduction

With regard to accented and dialectal speech recognition, a great deal of work has been done at various levels. Most of the dialect-specific automatic speech recognition (ASR) systems are concentrated on lexicon adaptation by capturing the pronunciation variations between standard speech and dialectal speech, and furthermore, characterizing these variation trends via a pronunciation lexicon [1-3]. Different from phone-level pronunciation modeling, the state-level pronunciation modeling is implemented to cover both the dialectal and the standard pronunciation characteristics [4, 5]. With regard to acoustic modeling, the adaptation techniques are most widely used through which dramatically significant improvement can usually be achieved [6, 7]. Some retraining mechanisms have also been proposed in which standard speech and dialectal/accented speech are pooled together [8]. Some researchers pay

Q. Huo et al. (Eds.): ISCSLP 2006, LNAI 4274, pp. 282 – 293, 2006.

attention to language adaptation for accented speakers [9]. Additionally, the decoder is adjusted to cope with the differences between standard speech and dialectal speech [2]. In practice, these approaches are always integrated together to achieve much better performance in dialectal/accented speech recognition.

As far as acoustic modeling for accented speech recognition is concerned, a couple of methods are usually used, including: 1) *Adaptation*. The acoustic models trained with standard speech are transformed into accent-specific ones with a certain amount of accented speech by means of adaptation. The adaptation method has been applied by many researchers to the accented speech recognition with good results. However, while the pronunciations in the target accent/dialect being primarily considered, those in the original accent/dialect cannot be sufficiently covered simultaneously at acoustic level. 2) *Retraining*. It is the most straightforward approach that pools accented training data with standard data so as to retrain the acoustic models using combined data. In [10], it is shown that by simply pooling 34 hours of standard data with 52 minutes of accented data the word error rate can be reduced from 49.3% to 42.7%. Although significant improvement was achieved with a small amount of accented data for "*pooled*" training, an obvious disadvantage was that the retraining was dramatically time-consuming. 3) *Combination of acoustic modeling with state-level pronunciation modeling* [4, 5]. In [4], state-level pronunciation modeling was integrated with acoustic modeling to better characterize the phone changes in which a syllable error rate (SER) reduction of approximately 2.39% was achieved for spontaneous speech recognition. The problem is that a large amount of accented speech data is needed and that the proposed method is sometimes too complicated to be readily applied. 4) *Dialect detection* [11, 12]. It is often used as a front-end in state-of-the-art ASR systems. In this method, dialect-specific recognizers have to be built for each dialect or sub-dialect, which also needs a large amount of dialectal data, and the performance, relies heavily on the outcome of dialect detection.

In China, *Putonghua* (or standard Chinese) is the official language through which people from different regions can be mutually understood. In addition to *Putonghua*, there are other 8 major dialects, which can be detailedly divided into over 40 sub-dialects [6] or over 1,000 sub-sub-dialects [13]. *Putonghua* spoken by most Chinese people is usually influenced by their native dialect more or less. In this paper, we refer to *Putonghua* influenced by a certain Chinese dialect as *dialectal Chinese*. One of our motivations here is to build a robust recognizer for a certain dialectal Chinese based on the handy *Putonghua* model with a small amount of dialectal speech data (less than one hour). To build a robust and practical dialectal Chinese-specific recognizer, the following four requirements should be met: 1) the modeling method as simple as possible, which is a prerequisite for fast deployment of ASR systems; 2) only a small amount of dialectal speech data needed. In China, there are so many dialects that it is impossible to collect a large amount of speech data for each dialectal Chinese due to some economical considerations; 3) good performance in dialectal speech recognition as well as standard speech recognition. Essentially, a dialect-specific recognizer is regarded as the extension of a *Putonghua* recognizer. It is natural that the better performance should be obtained for dialectal Chinese speech recognition without (or almost without) any performance degradation for *Putonghua* speech recognition; 4) a complementary or additive approach to the existing adaptation techniques. It is generally believed that adaptation is one of the most effective ways for speech

recognition of a dialectal Chinese of interest. Hopefully, the proposed modeling method can be used as a complement for the adaptation techniques in order to further improve the performance.

In order to reach the goal mentioned above, a novel, simple but effective acoustic modeling method is proposed in this paper, named as *state-dependent phoneme-based model merging* (SDPBMM) method. In SDPBMM, based on a same phoneme, the state-level parameters from a context-dependent *Putonghua* HMM and its phoneme-related context-independent dialectal HMM are merged according to a certain criterion. The idea comes from the assumption that the HMM from standard speech can "*borrow*" some information from its corresponding HMM in the target dialectal speech in order to reduce the differences between the dialectal speech and the standard speech. To a great extent, the newly-merged HMM can cover both dialectal and standard speech acoustically. In this paper, with only 40-minute Shanghai-dialectal speech data adopted, a cost-effective acoustic model for the target dialectal Chinese can be built from the *Putonghua* recognizer using SDPBMM method. It is experimentally shown that SDPBMM is able to meet the foresaid four requirements.

As a side effect of SDPBMM, the number of Gaussian mixtures within the merged HMMs is increased definitely, we regard it as a Gaussian mixtures expansion problem. To deal with it, an acoustic distance measure, named *pseudo-divergence based distance measure* (PDBDM), is proposed based on the difference measurement of Gaussian mixture models, and then implemented under the assumption that the similarity between two states can be measured by an acoustic distance between them. As a result, PDBDM can differentiate the states that need model merging from those that do not need merging in SDPBMM. More importantly, PDBDM can downsize the parameter scale of HMMs almost without causing any performance degradation on dialectal Chinese speech.

The remainder of this paper is organized as follows. The basic ideas of the SDPBMM will be described comprehensively in Section 2. In Section 3, a merging criterion, namely PDBDM, will be introduced which is to reduce the parameter scale in the SDPBMM-based HMMs. A series of experiments designed to evaluate the effectiveness of the proposed methods as well as the experimental results will be presented in Section 4. Finally, conclusions are drawn and future work is suggested in Section 5.

2 State-Dependent Phoneme-Based Model Merging

2.1 Description and Formulation of SDPBMM

In [4], a state-level pronunciation modeling method, the *partial change phone models,* was proposed, which could cover both the base form pronunciation and the surface form pronunciation simultaneously. The actually realized pronunciations except for the canonical pronunciation were merged with the pre-trained base form-based acoustic models in terms of the *acoustic model reconstruction*. Inspired by the idea, we make an attempt to take the standard pronunciation and dialectal pronunciation

into consideration in acoustic modeling. In SDPBMM, the context-dependent HMMs for standard Chinese are merged with their phoneme-related context-independent HMMs for dialectal Chinese at state level. In other words, the "*correct*" (base form) pronunciations in dialectal speech are involved in the merging instead of "*wrong*" (surface form) pronunciations adopted in [4, 5]. Due to the data sparseness issue, only monophone HMMs for dialectal Chinese are considered in SDPBMM. Compared with the *acoustic model reconstruction* based on triphone HMMs, a remarkably small amount of dialectal data is needed to build monophone HMMs and no further training process is necessary.

Most the state-of-the-art ASR systems tend to use context-dependent triphone HMMs to achieve a higher accuracy. In order to reduce the model complexity, downsize the redundant Gaussian components and re-estimate the unseen triphones in training data, the decision tree based state tying method is commonly used [14]. The states from some triphones with the same central phoneme are presented by a decision tree in which the tied states are presented by a leaf node. The idea is illustrated in Figure 1. In the left part of Figure 1, all the second states of the *an*-centered triphones are presented by a decision tree. In addition, both a state of monophone from dialectal speech and a state of triphone from standard speech are composed of multiple Gaussian mixtures. To accomplish the merging, the second state from the dialectal monophone *an* is merged with the leaf nodes of *an*-centered decision tree, *i.e.* the tied states. The merging process is depicted in the right part of Figure 1. The merging takes place between a monophone from dialectal speech and a triphone from standard speech whose central phoneme is the same as the monophone at the state level. As a result, a merged tied-state consists of multiple Gaussian mixtures from both the state of standard triphone HMM and its corresponding state of dialectal monophone HMM, as denoted by black solid curves and red dotted curves in Figure 1, respectively.

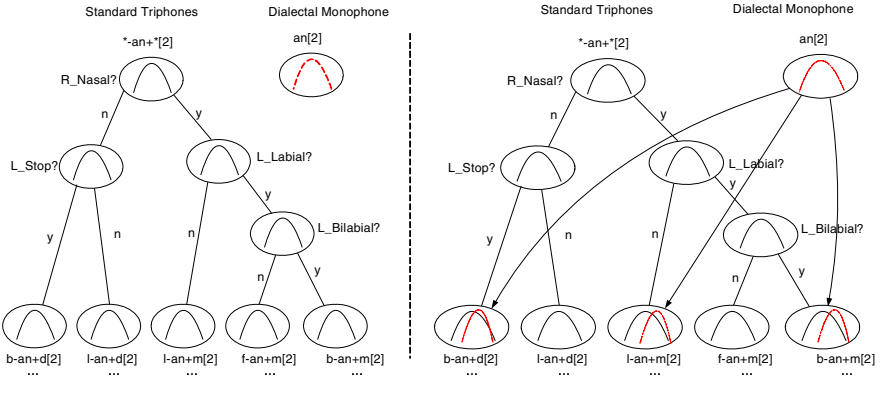

Fig. 1. The topology before and after the application of SDPBMM

Let x, s, and d be an input vector, a state from standard speech, and a state from dialectal speech, respectively, the original probability density function for continuous density HMM $P(x|s)$ is

$$P(x|s) = \sum_{k=1}^{K} w_{sk} N\left(x; \mu_{sk}; \Sigma_{sk}\right) \ , \tag{1}$$

where w_{sk} is the mixture weight of k-th mixture component of state s, K is the total number of Gaussian mixtures in state s. For simplification, $N_{sk}(\cdot)$ will be used to denote $N(x;\mu_{sk};\Sigma_{sk})$ of state s hereinafter.

Let $P'(x|s)$ be the revised output distribution of a merged state after applying SDPBMM, it can be represented as

$$P'(x|s) = \lambda P(x|s) + (1-\lambda) P(x|s,d) P(d|s) \ , \tag{2}$$

where λ is a linear interpolating coefficient between the standard and the dialectal acoustic models and is usually determined experimentally, and $P(d|s)$ can be regarded as a kind of pronunciation modeling. Because the purpose here is to verify the effectiveness of SDPBMM, the pronunciation variations between standard pronunciation and dialectal pronunciation are not taken into consideration in this paper and therefore we set $P(d|s) \equiv 1$. Afterwards, Equation 2 can be further simplified as Equations 3 and 4.

$$P'(x|s) = \lambda P(x|s) + (1-\lambda) P(x|d) \ , \tag{3}$$

$$\begin{aligned} P'(x|s) &= \lambda \sum_{k=1}^{K} w_{sk} N_{sk}(\cdot) + (1-\lambda) \sum_{n=1}^{N} w_{dn} N_{dn}(\cdot) \\ &= \sum_{k=1}^{K} \lambda w_{sk} N_{sk}(\cdot) + \sum_{n=1}^{N} (1-\lambda) w_{dn} N_{dn}(\cdot) \\ &= \sum_{k=1}^{K} w'_{sk} N_{sk}(\cdot) + \sum_{n=1}^{N} w'_{dn} N_{dn}(\cdot) \ . \end{aligned} \tag{4}$$

Equation 3 is actually a kind of interpolation method [8]. In Equation 4, K and N are the numbers of Gaussian mixtures of state s from standard speech and state d from dialectal speech, respectively; nevertheless parameters K and N are not necessary to be equal to each other. Parameters w'_{sk} and w'_{dn} are new mixture weights in a merged state of SDPBMM, just as indicated in Equation 4, $w'_{sk} = \lambda w_{sk}$ and $w'_{dn} = (1-\lambda) w_{dn}$.

2.2 Analysis

In SDPBMM, it is very easy to build context-independent monophone HMMs via just a quite small amount of dialectal data, and the merging is performed based on a standard triphone decision tree at state level. The SDPBMM-based acoustic model does not need retraining, which will save time and efforts dramatically. In essence, the SDPBMM-based acoustic model is still a standard recognizer just with much more acoustic coverage on dialectal speech, and so it is expected to be able to achieve good performance for both dialectal speech and standard speech recognition.

3 Pseudo-divergence Based Distance Measure

With the application of SDPBMM, the acoustic coverage is enlarged so that the accuracy for dialectal speech recognition can be improved; however, the Gaussian mixtures in the merged states are definitely increased. The efficiency is lowered due to much time consumption during the decoding procedure. For example, when a standard state consisting of 14 Gaussians is merged with a dialectal state of 6 Gaussians, the number of Gaussian mixtures is increased by 43% and thereby the time consumption is increased by 56% if all standard states are involved in the merging process. This is a Gaussian mixtures expansion problem. To deal with it, a mechanism has to be proposed to tell the states that need merging from those that do not need merging. Intuitively, there exists a different level similarity among states of dialectal monophone and states of standard triphone. Presumably, some measures can be taken to evaluate the similarity which can act as a criterion to classify the states participating in merging process. In practice, the similarity can be measured by the distance between two states instead.

In HMMs, each state is represented by a probability distribution function (PDF) in terms of mixed Gaussian mixtures. Several approaches have been proposed to measure the distance between two HHM states. 1) The relative entropy or Kullback Leibler distance (KLD) [15], which can represent the distance comprehensively but accordingly the computation complexity will easily go beyond control with the increased dimension. 2) Extended KLD, which is a practical way to approximate the distance [16]. But it can not be used to deal well with mixed-mixture PDFs and great time consumption is required. 3) Parametric distance metric for mixture PDF [17], which can effectively measure the distance directly between PDFs with mixed mixtures from the model's parameter. Actually, this approach is an issue of linear programming and can be solved via simplex tableau. However, sometimes the optimal solution can not be obtained under some rigid constraints. In this paper, as a tradeoff between precision and efficiency, a distance measure, named *pseudo-divergence based distance measure* (PDBDM), which was initially used and implemented in speaker recognition [18], is modified here to act as the distance measure between a state of dialectal monophone HMM and a tied-state of standard triphone HMM.

3.1 Basic Idea of PDBDM

In this section, the basic idea of PDBDM is to be illustrated in detail. First, the *dispersion* between two HMM states is defined as

$$dispersion(A,B)=\sum_{i=1}^{M} w_{Ai}\left[\sum_{j=1}^{N} w_{Bj} d_{A,B}(i,j)\right], \tag{5}$$

where A and B are two HMM states, $d_{A,B}(i, j)$ is the distance between the i-th mixture from A and j-th mixture from B. M and N are the total numbers of Gaussian mixtures in A and B, respectively. Accordingly, the *self-dispersion* is

$$dispersion(A,A)=\sum_{i=1}^{M} w_{Ai}\left[\sum_{j=1}^{M} w_{Aj} d_{A,A}(i,j)\right]. \tag{6}$$

Then the *pseudo-divergence* between two HMM states is formulated as[1]:

$$pseudo\text{-}divergence\left(\lambda_A,\lambda_B\right)=\frac{dispersion(A,B)}{dispersion(A,A)}\ , \tag{7}$$

$$pseudo\text{-}divergence\left(\lambda_B,\lambda_A\right)=\frac{dispersion(B,A)}{dispersion(B,B)}\ . \tag{8}$$

Usually $pseudo\text{-}divergence\left(\lambda_A,\lambda_B\right)\neq pseudo\text{-}divergence\left(\lambda_B,\lambda_A\right)$. To minimize the statistical difference, the distance between two HMM states is redefined as

$$distance\left(\lambda_A,\lambda_B\right)=\frac{1}{2}\left(\frac{dispersion(A,B)}{dispersion(A,A)}+\frac{dispersion(B,A)}{dispersion(B,B)}\right)\ . \tag{9}$$

As for the distance between two single Gaussian mixtures, *i.e.* $d_{A,B}(i, j)$ in Equation 5, there are normally four options, the *Euclidean* distance measure, the *Mahalanobis* distance measure, the *weighted Mahalanobis* distance measure and the *Bhattachyaryya* distance measure. The Bhattachyaryya distance measure is adopted here because it is thought to be able to characterize the distance more precisely by taking the difference of covariance into account [17]. Given two Gaussian mixtures, $\lambda_1(\mu_1, \Sigma_1)$ and $\lambda_2(\mu_2, \Sigma_2)$, the Bhattachyaryya distance measure is defined as

$$d(\lambda_1,\lambda_2)=\frac{1}{8}(\mu_1-\mu_2)^T\left(\frac{\Sigma_1+\Sigma_2}{2}\right)^{-1}(\mu_1-\mu_2)+\frac{1}{2}\ln\frac{\left|(\Sigma_1+\Sigma_2)/2\right|}{\left|\Sigma_1\right|^{1/2}\left|\Sigma_2\right|^{1/2}}\ . \tag{10}$$

3.2 Combination of SDPBMM with PDBDM

A state from a dialectal monophone HMM and its corresponding state on a basis of the same phoneme from a standard triphone HMM form a pair for the calculation of distance. The distances of all pairs are computed using Equation 9. Subsequently, a certain percentage, *i.e.* 70% relative to the amount of pairs, is set as a *threshold* in the descending order of distance so that the pairs with a large distance have a higher priority to be chosen to participate in the merging. The idea is depicted in Equation 11.

$$\begin{cases} merging, & \text{distance}(d,s)\geq threshold \\ no\text{-}merging, & \text{distance}(d,s)<threshold \end{cases} \tag{11}$$

The application of PDBDM in SDPBMM is based on the assumption that the distance can be used to characterize the similarity between two states instead, but in a reverse sense that a smaller distance corresponds to a bigger similarity. If the distance between two states is small, it can be safely inferred that there is less variability between them and in which case no merging is necessary because the original state

[1] In the paper, the concept of *divergence* is not completely same as the classic definition of *divergence*, so *pseudo-divergence* is named.

from the standard speech has already covered the acoustic space sufficiently. As for the pairs with big distances, the merging is performed to cover both the standard and the dialectal speech acoustically. Notice the fact that in the right part of Figure 1, some states, for example *l-an+d*[2] and *f-an+m*[2], are not involved in the merging, which represents the purpose of PDBDM. It is expected that the scale of Gaussian mixtures can be downsized by PDBDM while no performance degradation takes place for dialectal speech recognition.

4 Experiments and Results

The Mandarin Broadcast News (MBN) database (Hub4NE), a read style standard Chinese speech corpus, was used to train the baseline system, the *Putonghua* recognizer. It contained about 30 hours of high quality wideband speech with detailed Chinese Iinital/Final (IF) transcriptions. The acoustic models of *Putonghua*-based baseline were tied-state cross-word standard tri-IF HMMs. Each tri-IF was modeled using a left-to-right non-skip 3-state continuous HMM, with 14 Gaussian mixtures per state. 39-dimensional MFCC coefficients with Δ and $\Delta\Delta$ were used as features with cepstral mean normalization [19]. The HMMs achieve good performance statistically upon which many research was carried out [12]. Additionally, 6 zero-Initials were added to the standard IF set to help improve the performance and make the modeling process consistent. Another database, namely Wu dialectal Chinese database (WDC) [20], contained 100 native Shanghai speakers, 50 males and 50 females. The speech data of WDC was recorded under a similar condition to that of MBN. The use of this database was to minimize the channel affect. The WDC was composed of the speech from medium and strong Shanghai-accented speakers. Further details on the database can be found in [20]. Adopted in the following experiments as the recognition lexicon were 406 toneless Chinese syllables.

Three data sets were selected from the WDC and MBN, one was the development training set, *Dev_WDC*, which consisted of about 40-minute Shanghai-dialectal Chinese speech by 10 speakers. The *Dev_WDC* was used to build 65 context-independent dialectal mono-IF HMMs for SDPBMM, each monophone HMMs was of the exactly same topology as that of standard tri-IFs except that there were 6 Gaussian mixtures per state. Another data set was *Test_WDC* composed of 20 speakers' speech from the WDC. The third data set was *Test_MBN* from MBN also used for testing. The three data sets were not overlapping with one another. The detailed information for the data sets used in the experiments is listed in Table 1. Initially, the MBN-based *Putonghua* HMMs achieved SERs of 30.5% and 49.8% on *Test_MBN* and *Test_WDC*, respectively; there was an absolute degradation of approximately 20% on the Shanghai-dialectal Chinese speech. An SER of 54.1% on *Test_WDC* was achieved by the dialectal mono-IF HMMs built upon *Dev_WDC*. Because acoustic modeling was our research focus no language models were used. Our experiments were performed at the syllable level and the SER reduction was used as a measure of the improvement. Besides, HTK [21] was used in the experiments.

Table 1. Detailed information for the development and test sets

Data set	Database	Details
Dev_WDC	WDC	10 speakers, 510 utterances, totally 40-minute speech
Test_WDC	WDC	20 speakers, 995 utterances, totally 60-mintue speech
Test_MBN	MBN	1,200 utterances, totally 80-minute speech

The linear coefficient in Equation 3 was determined experimentally and λ was set to 0.72. With the application of PDBDM, there were 70% of tied states from *Putonghua* tri-IFs involved in the SDPBMM. The recognition results on the dialectal test set, *Test_WDC*, are listed in Table 2. It can be seen that the SDPBMM can reduce the SER by 6.2% absolutely on dialectal speech with only 40-minute dialectal Chinese speech data. However the number of Gaussian mixtures was increased by approximately 43%. To deal specifically with the Gaussian mixtures expansion problem, the PDBDM was adopted with the expectation that no degradation is introduced. Thus, the number of Gaussian mixtures in *SDPBMM+PDBDM* was decreased by 30% with a slight SER increase of 0.3% absolutely. Compared with the baseline, an absolute SER reduction of 5.9% was still achieved by the *SDPBMM+PDBDM*. It is shown that PDBDM can downsize the parameter scale without significant performance degradation. In the following experiments, *SDPBMM+PDBDM* was used as the default SDPBMM-based acoustic modeling.

Table 2. The results for *Putonghua*, SDPBMM, and SDPBMM in conjunction with PDBDM on *Test_WDC*

	Putonghua	SDPBMM	SDPBMM+PDBDM
States	3,230	3,230	3,230
Gaussians	45,220	64,600	58,786
Tri-IFs	7,411	7,411	7,411
SER	49.8%	43.6%	43.9%

4.1 Comparison Conditioned on Same Amount of Gaussian Mixtures

As for SDPBMM-based acoustic model, it is naturally assumed that the improvement in dialectal speech recognition may result from the increase of Gaussian mixtures in the merged states. Compared with the *Putonghua* HMMs with 14 Gaussian mixtures per state, on average, there were 18.2 mixtures per state in SDPBMM-based HMMs. To make a fair comparison, another *Putonghua* acoustic model with 18 Gaussian mixtures per state was generated which had approximately equal parameter scale as that of the SDPBMM. The SER on *Test_WDC* was decreased from 49.8% to 49.1% compared with the baseline, but there still existed an SER gap of 5.2% absolutely in comparison with the SDPBMM. It is shown that increasing the parameter scale solely can not achieve significant improvement in dialectal speech recognition.

4.2 Evaluation on Standard Speech Recognition

The effectiveness of SDPBMM-based acoustic model on standard speech recognition can be seen from the results listed in Table 3 with *Test_MBN* taken as the test set. It is

shown that as expected, the SDPBMM can achieve a slightly higher SER (an absolute 0.6% higher) on standard speech than the *Putonghua* acoustic model. It could be concluded that the SDPBMM can achieve significant improvement in dialectal speech recognition without significant degradation in standard speech recognition.

Table 3. The results for *Putonghua* and SDPBMM on *Test_MBN*

	Putonghua	SDPBMM
SER	30.5%	31.1%

4.3 Integration with Adaptation

Adaptation is one of the most effective ways for dialectal speech recognition. Most widely used adaptation techniques include the maximum linear likelihood regression (MLLR) and the maximum *a posteriori* adaptation (MAP) methods [19]. For comparison, the adaptation was performed with exactly the same amount of dialectal speech data as in the experiment regarding SDPBMM. Considering that MLLR is much beneficial when there is only a small amount of adaptation data available [7], we adopted MLLR for model adaptation. The MLLR adaptation was performed based on the *Putonghua* acoustic model, denoted as *MLLR*, in which all the standard tri-IFs were classified into 65 classes, and mean update was performed in transformation matrix. Note that, *Dev_WDC* was also used as the adaptation data in MLLR adaptation. As a result, an SER of 44.1% was achieved on *Test_WDC* which was still slightly higher than the SER of 43.9% by SDPBMM with exactly the same data set. The results are listed in columns *SDPBMM* and *MLLR* in Figure 2, respectively. It is shown that compared with MLLR, SDPBMM can achieve a comparable performance on dialectal speech recognition with only a small amount of dialectal data available.

In addition, it is assumed that SDPBMM primarily concentrates on addressing the issues of the phonetic mismatch between the dialectal speech and the standard speech. As a matter of fact, the adaptation can be a good solution to channel mismatch. Therefore it is expected that the SDPBMM in combination with a certain adaptation method can have the potential to further improve the performance on dialectal speech recognition. To verify the assumption, another development data set of Shanghai-dialectal Chinese, *Dev_WDC*1, was selected from WDC database, which consisted of 410 utterances by 10 speakers (approximately 30 minutes). By using *Dev_WDC* and *Dev_WDC*1, two new acoustic models were built, namely *SDPBMM+MLLR* and *MLLR+SDPBMM*, where the order in the names means the order that the components were performed. In *SDPBMM+MLLR*, the SDPBMM was performed using *Dev_WDC* based on *Putonghua* HMMs followed by the MLLR adaptation using *Dev_WDC*1; and vice versa. The results are also listed in Figure 2. From the figure, it can be clearly seen that in combination with the MLLR adaptation, another two absolute SER reductions of 1.9% and 1.8% on dialectal Chinese speech can be further achieved by *SDPBMM+MLLR* and *MLLR+SDPBMM*, respectively. The results correspond to columns *SDPBMM+MLLR* and *MLLR+SDPBMM* in Figure 2, respectively. Another phenomenon is that *SDPBMM+MLLR* and *MLLR+SDPBMM* achieved approximately an equal SER, which is to say, the SDPBMM and MLLR can collaborate perfectly irregardless of the application order. In conclusion, SDPBMM and MLLR are additive and exchangeable algebraically.

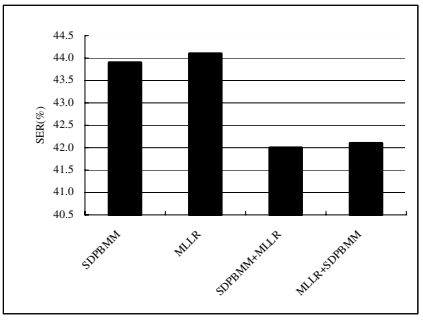

Fig. 2. Comparison with MLLR adaptation on *Test_WDC* and integration with MLLR adaptation on *Test_WDC* and *Test_WDC*1

5 Conclusions and Future Work

In the paper, a novel, simple but effective acoustic modeling method for dialectal Chinese speech recognition, SDPBMM, is proposed. Though it will introduce a Gaussian mixtures expansion problem, a corresponding PDBDM acting as a merging criterion is proposed to be integrated into SDPBMM, which can result in no significant degradation for dialectal Chinese speech recognition. From a series of experiments, it can be concluded that the SDPBMM has the advantages: 1) It is simple but practical for acoustic modeling when there is quite a small amount dialectal speech data available; 2) It can make a significant performance improvement for dialectal speech recognition; 3) It can have good performance for both standard and dialectal speech recognition; 4) It can achieve comparable performance to adaptation with only a small amount dialectal speech data available; 5) It is additive to adaptation, that is to say, the application of SDPBMM and adaptation in any order can further improve the performance for dialectal speech recognition. In a word, the SDPBMM is one of the most effective acoustic modeling methods for read-style dialectal Chinese speech recognition. In this paper, the experiments were done on Shanghai-dialectal Chinese, but no dialect-specific prior knowledge is incorporated in SDPBMM, thus, this method can be easily generalized to other dialectal Chinese.

Another issue is that the experiments in this paper were based on read speech. In our next step the research on spontaneous speech will be carried out where pronunciation modeling [22] should be taken into account. It is believed that the use of pronunciation modeling can help build much precise acoustic model to better characterize pronunciation variations between dialectal Chinese and *Putonghua*, not only for spontaneous speech but also for read speech.

References

1. Goronzy, S., Kompe, R., Rapp, S.: Generating Non-Native Pronunciation Variants for Lexicon Adaptation. Speech Communication, Vol. 42(1):109-123, 2004
2. Huang, C., Chen, T., Chang, E.: Accent Issue in Large Vocabulary Continuous Speech Recognition. International Journal of Speech Technology, 7: 141-153, 2004

3. Tjalve, M., Huckvale, M.: Pronunciation Variation Modeling using Accent Features. Proc. Interspeech, 2005, Lisbon
4. Liu, Y., Fung, P.: Pronunciation Modeling for Spontaneous Mandarin Speech Recognition. International Journal of Speech Technology, 7:155-172, 2004
5. Saraclar, M., Nock, H., Khudanpur, S.: Pronunciation Modeling by Sharing Gaussian Densities across Phonetic Models. Computer Speech and Language, 14:137–160, 2000
6. Li, J., Zheng, T.-F., Byrne, W., Jurafsky, D.: A Dialectal Chinese Speech Recognition Framework. Journal of Computer Science and Technology, 21(1): 106-115, Jan. 2006
7. Diakoloukas,V., Digalakis, V., Neumeyer, L., Kaja, J.: Development of Dialect-Specific Speech Recognizers Using Adaptation Methods. IEEE ICASSP, 2:1455, 1997.
8. Tomokiyo, L.-M.: Recognizing Non-Native Speech: Characterizing and Adapting to Non-Native Usage in LVCSR. PhD Thesis, Carnegie Mellon University, 2001.
9. Gao, J.-F., Goodman, J., Li, M.-J., Lee, K.-F.: Toward a Unified Approach to Statistical Language Modeling for Chinese. ACM Transactions on Asian Language Information Processing, 1(1): 3- 33, March 2002
10. Wang, Z.-R., Schultz, T., Waibel, A.: Comparison of Acoustic Model Adaptation Techniques on Non-native Speech. IEEE ICASSP, 540–543, 2003
11. Zheng,Y.-L., Sproat, R., Gu, L. *et al*.: Accent Detection and Speech Recognition for Shanghai-Accented Mandarin", Interspeech 2005, Lisbon
12. Sproat, R., Zheng, T.-F., Gu, L., Jurafsky, D., Shanfran, I., Li, J., Zheng, Y.-L., Zhou, H., Su, Y., Tsakalidis, S., Bramsen, P., Kirsch, D.: Dialectal Chinese Speech Recognition: Final Technical Report. 2004, http://www.clsp.jhu.edu/ws2004/
13. Li, A.-J., Wang, X.: A Contrastive Investigation of Standard Mandarin and Accented Mandarin, EuroSpeech, 2003, Geneva
14. Hwang, M.-Y., Huang, X.-D., Alleva, F.-A.: Predicting Unseen Triphones with Senones, IEEE Transaction on Speech and Audio Processing, 4(6):412-419, 1996
15. Cover, T.-M., Thomas, J.-A.: Elements of Information Theory, John Wiley & Sons, 1991
16. Liu, Z., Huang, Q.: A New Distance Measure for Probability Distribution Function of Mixture Types. Proc. ICASSP, 1345-1348, 2000
17. Liu, Y., Fung, P.: Acoustic and Phonetic Confusions in Accented Speech Recognition. Proc. INTERSPEECH, 3033-3036, 2005
18. Xuan, P., Wang, B.-X.: Speaker Clustering via Distance Measurement of Gaussian Mixtures Models. Journal of Computer Engineering and Technology, May, 2005
19. Huang, X.-D. Acero, A., Hon, S.-W.: Spoken Language Processing, Prentice Hall, 2001
20. Li, J., Zheng, F., Xiong, Z.-Y. Wu, W.-H.: Construction of Large-Scale Shanghai Putonghua Speech Corpus for Chinese Speech Recognition, Oriental-COCOSDA, 62-69, October, 2003, Singapore
21. Young, S., Evermann,G., Hain,T. *et al.*: The HTK Book (for HTK Version 3.2.1). Cambridge University, Cambridge, 2002. http://htk.eng.cam.ac.uk/
22. Zheng, F., Song, Z.-J., Fung, P., Byrne, W.: Mandarin Pronunciation Modeling Based on CASS Corpus, Journal of Computer Science and Technology, 17(3): 249-263, May 2002