Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 50 (2008) 605–615
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

State-dependent phoneme-based model merging for dialectal Chinese speech recognition ☆

Linquan Liu, Thomas Fang Zheng *, Wenhu Wu

Center for Speech and Language Technologies, Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing 100084, China

Received 28 November 2006; received in revised form 6 March 2008; accepted 24 April 2008

Abstract

This paper discusses and evaluates a novel but simple and effective acoustic modeling method called "state-dependent phoneme-based model merging (SDPBMM)", used to build dialectal Chinese speech recognizer from a small amount of dialectal Chinese speech. In SDPBMM, state-level pronunciation modeling is done by merging a *tied-state* of *standard triphones* with a *state* of *dialectal monophone*(*s*). In state-level pronunciation modeling, which acts as the merging criterion for SDPBMM, sparseness arises due to limited data set. To overcome this problem, a distance-based pronunciation modeling approach is also proposed. With a 40-min Shanghai-dialectal Chinese speech data, SDPBMM achieves a significant absolute syllable error rate (SER) reduction of approximately 7.1% (and a relative SER reduction of 14.3%) for Shanghai-dialectal Chinese, without performance degradation for standard Chinese. It is experimentally shown that SDPBMM outperforms Maximum Likelihood Linear Regression (MLLR) adaptation and the Pooled Retraining methods by 1.4% and 5.3%, respectively, in terms of SER reduction. Also, when combined with MLLR adaptation, an absolute SER reduction of 1.4% can further be achieved by SDPBMM.

Keywords: Speech recognition; Dialectal Chinese; State-dependent phoneme-based model merging; Acoustic modeling; Pronunciation modeling; Acoustic model distance measure; A small amount of data

1. Introduction

Accent is one of the most challenging issues in current automatic speech recognition (ASR) systems. Dialectal speech is a variant of a language spoken by people living in a certain dialect region, and hence is of similar or identical accent with some distinct and unique regional characteristics. For the past few years, efforts have been made to improve accented speech recognition accuracy. As a result, some promising results are achieved. Usually, research on accented speech recognition is being carried out on the following aspects.

(1) Pronunciation modeling. Pronunciation lexicon has been the primary focus of dialectal speech recognition (Goronzy et al., 2004; Huang et al., 2004; Tjalve and Huckvale, 2005). Efforts have also been made on state-level pronunciation modeling for accented speech recognition (Liu and Fung, 2004; Saraclar et al., 2000).
(2) Retraining (Tomokiyo, 2001; Wang et al., 2003). Retraining could be done to build a robust ASR system by pooling standard and dialectal speech together; or using only dialectal speech to build an accented speech recognition system for that dialect.

☆ This paper is based upon a study partially supported by the US National Science Foundation (NSF) under Grant No. 0121285 and Sony Computer Entertainment Inc. (SCEI). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of NSF or SCEI.
* Corresponding author. Tel./fax: +86 10 62796393.
E-mail addresses: liulq@cslt.riit.tsinghua.edu.cn (L.-Q. Liu), fzheng@tsinghua.edu.cn (T.F. Zheng), wuwh@tsinghua.edu.cn (W.-H. Wu).

0167-6393/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2008.04.006

(3) Adaptation (Li et al., 2006; Diakoloukas et al., 1997). Adaptation is one of the most effective techniques to improve performance in accented speech recognition. Adaptation could be performed on acoustic model as well as language model. MLLR and maximum *a posteriori* (MAP) are commonly used to adapt acoustic models, whereas cache language models, topic-adaptive models and maximum entropy models are choices for language models (Gao et al., 2002; Gruhn et al., 2004).
(4) Decoder tuning (Huang et al., 2004). Modifications are made to the decoder to better characterize accented speech.
(5) Accent or dialect classification (Zheng et al., 2005; Sproat et al., 2004). Classification process is performed before speech recognition. It is usually employed as a front-end for ASR systems.

In practice, the aforementioned approaches are combined in various ways to achieve better performance in accented speech recognition.

Putonghua (or standard Chinese) is the official language of China. Often influenced by local dialects, it is spoken with a vast variation, but still intelligible to people around China. Besides *Putonghua*, there are other 8 major dialects, which could be divided into over 40 sub-dialects (Li et al., 2006), further divided into some 1000 descendant dialects (Li and Wang, 2003). This is why *Putonghua*, spoken by most Chinese people, is fairly influenced by their local dialects. In this paper, we will refer to *Putonghua* spoken by people living in a certain region and influenced by its corresponding Chinese dialect as *dialectal Chinese*. In contrast to accented speech recognition which primarily focuses on phonetic differences, dialectal speech recognition focuses not only on phonetic differences but also on syntax and semantic differences. Much details about the concepts of dialectal and accented speech recognition could be found in (Sproat et al., 2004; Li et al., 2006). As an extended research of Sproat et al. (2004) and Li et al. (2006), acoustic modeling for dialectal Chinese speech recognition is a major concern in our research.

In general, it is impractical to collect a large amount of data to build a recognizer for each kind of dialectal Chinese, owing to its diversity. Thus, one of our motivations is to build a robust recognizer for dialectal Chinese primarily based on *Putonghua*. However, it would incorporate dialectal Chinese by including a small amount of dialectal Chinese speech data (less than 1 h). Little literature has yet been available that focuses on using small amount of accented data when building an accent-friendly ASR system. Also, none of the state-of-the-art ASR systems has been known to achieve good performance for both dialectal speech and standard speech recognition simultaneously.

Attempting to address these issues, we propose a novel, simple and effective acoustic modeling approach – state-dependent phoneme-based model merging (SDPBMM). Here, Gaussian mixtures from context-dependent *Putonghua* triphone HMM and its phoneme-related context-independent dialectal monophone HMM are merged at state-level according to both the pronunciation variation probability and the interpolation coefficient between them. This idea comes from the assumption that an HMM from standard speech can "borrow" some information from its corresponding HMM in the target dialectal speech in order to narrow the gap between standard speech and dialectal speech. To a great extent, the newly-merged HMM can cover both the standard and the dialectal speech acoustically. In this paper, with only 40-min Shanghai-dialectal speech data adopted, a cost-effective acoustic model for the Shanghai-dialectal Chinese is built from a *Putonghua* recognizer using SDPBMM. It is experimentally shown that SDPBMM has many advantages in dialectal speech recognition.

State-level pronunciation modeling, which acts as a merging criterion, plays an important role in SDPBMM. Therefore, acquiring pronunciation variations and precisely calculating the variation probability, particularly based on a small amount of dialectal data, is an inevitable challenge. Naturally, how to deal with data sparseness becomes a key issue for pronunciation modeling. To address this issue, a distance-based pronunciation modeling approach is proposed in this paper.

This paper is organized as follows. In Section 2, a brief overview of state-of-the-art research on acoustic modeling for dialectal speech recognition is given, including those which became the inspiration and motivation for our proposal. The basic idea of SDPBMM is explained in Section 3. Distance-based pronunciation modeling with a small amount of dialectal speech data is discussed in Section 4. A series of experiments designed to evaluate effectiveness of the proposed methods as well as some experimental results are presented in Section 5. Conclusions are drawn and future work on dialectal speech recognition is outlined in Section 6.

2. Related work and objective for dialectal speech recognition

Although there exist differences between them, accented and dialectal speech recognition own many common phonetic characteristics, hence some acoustic modeling approaches can be shared between the two of them.

2.1. Latest research on acoustic modeling for accented speech recognition

Most of the work done on acoustic modeling for dialectal/accented speech recognition has primarily focused on the following aspects.

Integration of pronunciation modeling with acoustic modeling: Due to the fact that dialectal Chinese has similar pronunciation tendency or pronunciation habit among the speakers from a dialect region, it is a good candidate for pronunciation modeling to deal with pronunciation

variations caused by a local dialect. Most of the state-of-the-art ASR systems focus on lexicon adaptation. In this method an accent-specific pronunciation lexicon, based on data-driven or knowledge-based pronunciation modeling, is generated. Many researchers have made successful attempts to perform pronunciation modeling at sub-lexical level; others have integrated pronunciation variability into acoustic modeling for accented speech recognition. In (Liu and Fung, 2004), the state-level pronunciation modeling is integrated with acoustic modeling for better characterization of sound change, where an absolute SER reduction of 2.39% was achieved for spontaneous speech recognition. In (Oh et al., 2007), acoustic modeling adaptation based on pronunciation variability was performed; as a result of which word error rate (WER) was decreased from 39.2% to 33.2% on Korean-accented English.

Retraining: With regard to acoustic modeling for the dialectal speech recognition, pooling standard speech and dialectal speech for training is the most straightforward method. It is shown in (Wang et al., 2003) that by simply pooling 34 h of standard data with 52 min of accented data, WER can be reduced from 49.3% to 42.7%. Another similar approach is to train models on standard speech, and then perform a few forward–backward iterations with accented speech. In (Tomokiyo, 2001), acoustic models were built on native English first, and then two additional forward–backward iterations were performed using a 3-h Japanese-accented English dataset. Consequently, WER dropped from 63.0% to 48.0%. Acoustic model interpolation was also employed to smooth difference between standard speech and accented speech. In (Livescu, 1999), it is shown that interpolating native and non-native retrained acoustic models lead to a relative WER reduction of 8.1% on a non-native test set.

Adaptation: Given a small amount of dialectal data available, adaptation method is widely used for dialectal speech recognition. The MLLR adaptation is often employed when adaptation data is insufficient. In (Diakoloukas et al., 1997), speech data of 500 utterances from 10 speakers was used; reducing WER from 63.0% to 54.2%. MLLR adaptation based on a small amount of accented speech is also studied (He and Zhao, 2003; Xu and Yan, 2004). When sufficient training data becomes available, MAP usually outperforms MLLR because of much precise update of each Gaussian transformation (Myrvoll, 2003). Some researchers combined two adaptation methods for further improvement in performance. They used MLLR transformed means as *priors* for MAP adaptation (Sproat et al., 2004). The literature review leads us to believe that when small amount of dialectal data is available, MLLR is a fairly effective approach for dialectal speech recognition.

Dialect classification: Dialect classification often acts as a front-end for dialectal speech recognition. In such ASR systems, a dialect or sub-dialect is first identified by a dialect classification component. Word-based and phone-based dialect classification approaches are most commonly used (Huang and Hansen, 2005; Angkititrakul and Hansen, 2006). In (Chen et al., 2001), it is shown that 3–5 utterances were sufficient to recognize a speaker's dialect. In (Sproat et al., 2004; Li and Wang, 2003), Shanghai-dialectal Chinese was classified into several categories according to the speakers' accent levels. Hence, at the back-end in such ASR systems, each dialect or sub-dialect class corresponds to accent-specific acoustic models, lexica, and even language models.

2.2. Robust dialectal chinese speech recognizer

In China, *Putonghua* is spoken quite differently at different regions due to influence of local dialects. However, a lot of common characteristics are shared among them, so the various kinds of dialectal Chinese are regarded as variations of *Putonghua*. It can be reasonably assumed that a *Putonghua* recognizer is taken as a benchmark for recognizing dialectal Chinese. Dialectal Chinese also possesses many unique characteristics introduced by its corresponding local dialect. It is expected that in combination with a relatively small amount of dialectal speech, a *Putonghua* recognizer can be tuned to improve the performance in dialectal Chinese speech recognition without affecting the performance in *Putonghua* speech recognition. The basic framework is illustrated in Fig. 1. To achieve the goal, some effective measures should be taken to deal with the differences between *Putonghua* and a certain dialectal Chinese.

There exist several disadvantages in the state-of-the-art ASR systems for dialectal speech recognition. Retraining is quite time-consuming and a large amount of dialectal data is often required. Pronunciation modeling at lexical level is good at modeling phone changes, but gives poor

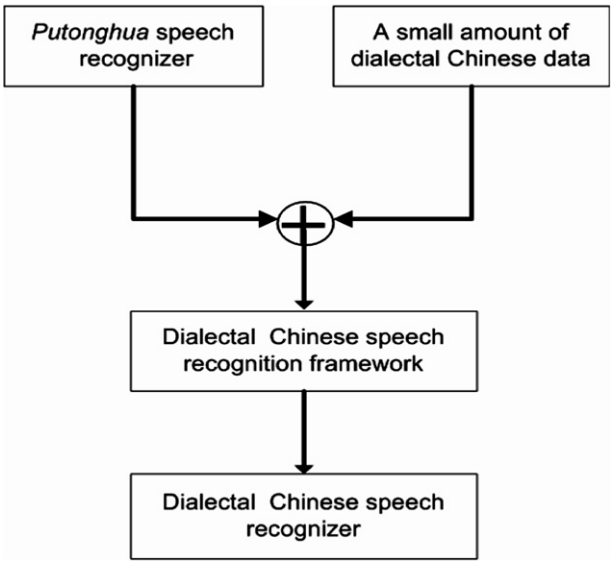

Fig. 1. Framework for dialectal Chinese speech recognizer.

results for sound changes in dialectal speech due to ambiguous pronunciation transformation. Alternatively, pronunciation modeling at state-level offers a good solution to sound changes for dialectal speech recognition; it is usually combined with acoustic modeling or adaptation. However, sometimes, the model sharing strategy is quite complicated, for example, in (Liu and Fung, 2004; Fung and Liu, 2005), some accent-specific units and their corresponding decision trees were generated during acoustic model reconstruction. Also, to obtain context-dependent HMMs for these accent-specific units, a larger amount of data is required. In (Oh et al., 2007), retraining was performed in combination with six pronunciation variants in Korean-accented English. In doing so, extra time consumption for retraining was needed. Furthermore, degradation on standard speech may result. Adaptation is very effective in recognition of dialectal speech. Unfortunately, it always transforms a standard model into a dialect-specific one, due to which good performance for standard speech and dialectal speech simultaneously cannot be guaranteed. Generally speaking, dialect classification is a good choice when dialectal speech is sufficient to build a dialect-specific recognizer. However, further classification within a specific dialectal Chinese is likely to suffer from data sparseness.

Keeping our objective in mind, and having studied previous related work, we came to a conclusion that a good approach to building an effective dialectal Chinese recognizer should meet the following requirements: (1) modeling method should be as simple as possible. This is a prerequisite for fast deployment of ASR systems; (2) only a small amount of dialectal speech data should be required keeping in view the dialect diversity and economic considerations; (3) good performance should be offered in both dialectal speech recognition and standard speech recognition. Since a dialectal Chinese recognizer is regarded as the extension and variation of a *Putonghua* recognizer, better performance should be obtained for dialectal Chinese speech recognition with almost no performance degradation for *Putonghua* speech recognition; (4) the new approach should be a complementary or additive approach to the existing techniques. In other words, the proposed method must not have an adverse effect on any methods, when combined with it. For instance, the proposed method should not antagonize adaptation methods or language modeling.

Attempting to meet the requirements mentioned above, we propose *state-dependent phoneme-based model merging* (SDPBMM) method in acoustic modeling for dialectal speech recognition, which will be explained in the following section.

3. The SDPBMM method

In (Liu and Fung, 2004), a state-level pronunciation modeling method, *Partial Change Phone Models*, was proposed. It can cover both canonical form pronunciation and surface form pronunciation simultaneously. The actual pronunciations except the canonical pronunciations are merged with the pre-trained, canonical form-based acoustic models, using *acoustic model reconstruction* method. Inspired by the idea, we attempted to incorporate both standard and dialectal pronunciations in acoustic modeling. That is, merging Gaussian mixtures from a context-independent monophone HMM for dialectal Chinese into their phoneme-related context-dependent triphone HMMs for *Putonghua* at state level, which is referred to as SDPBMM. The term *state-dependent* means that indices for states are identical in each HMM in the merging process. Apparently, only monophone HMMs for dialectal Chinese are involved in SDPBMM. In contrast to triphone HMM, monophone HMM not only requires significantly less observations, but also has fewer Gaussian mixtures per state; and hence the particularly crucial issue of data sparseness, which appears when limited dialectal speech data is available, can be mitigated. Compared with the *acoustic model reconstruction* based on triphone HMMs, a notably less dialectal speech data is required to build monophone HMMs.

Most of the state-of-the-art ASR systems tend to use context-dependent triphone HMMs to build robust acoustic models. In order to reduce the model complexity, trimming down the redundant Gaussian Mixtures and estimating unseen triphones in training data, decision tree-based state tying method is commonly used (Hwang et al., 1996). Correspondingly, SDPBMM is implemented within a decision tree. The states from triphone HMMs with the same central phoneme are represented by a decision tree in which the tied states are represented by a leaf node. That is, multiple states of triphone HMMs with the same centre-phone are treated the same in the decision tree. The idea is illustrated in Fig. 2. In Fig. 2, taking a Chinese Initial *an* as an example, the 2nd states of the *an*-centered triphones are represented by a decision tree. Both a state of monophone from dialectal speech and a tied-state of

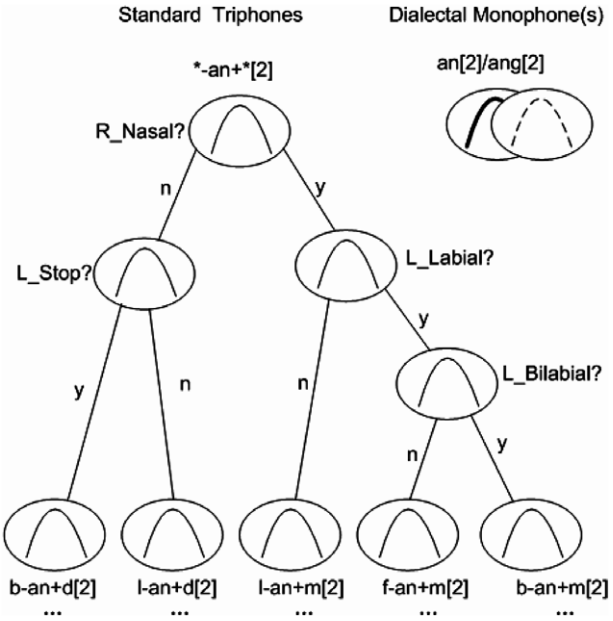

Fig. 2. Original topology before SDPBMM.

triphone from standard speech are composed of multiple Gaussian mixtures. To accomplish the merging, Gaussian mixtures within the 2nd states from the dialectal monophone *an* and its pronunciation variants *ang* are merged into the leaf nodes of *an*-centered decision tree, *i.e.* the tied states of standard triphone. In that case, whether a pronunciation variant *ang* for Initial *an* is involved in the merging or not is determined by pronunciation modeling which will be introduced in Section 4. In some sense, initial *an* is compulsory and its variant *ang* is optional in SDPBMM. The merging process is depicted in Fig. 3. The merging takes place at state level between a monophone together with its pronunciation variants from dialectal speech and a triphone from standard speech whose central phoneme is the same as the dialectal monophone. 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 phoneme-related dialectal monophone HMM(s), as denoted by thin solid curves, thick solid curves and thin dotted curves in Fig. 3, respectively.

Theoretically, SDPBMM is formulated as follows. Let x and s_i be an input vector and the ith state in a HMM, respectively. The original probability density function (PDF) for continuous HMM $p(x|s_i)$ is

$$p(x|s_i) = \sum_{k=1}^{K} w_{ik} N(x; \mu_{ik}; \Sigma_{ik}), \tag{1}$$

where w_{ik} is the mixture weight of the kth mixture component of state i, K is the total number of Gaussian mixtures in state i. For simplification, $N(x; \mu_{ik}; \Sigma_{ik})$ will hereinafter be denoted as $N_{ik}(\cdot)$.

Let $p'(x|s_i)$ be the revised output PDF of a merged state i after applying SDPBMM, it can be represented as

$$p'(x|s_i) = \lambda p(x|s_i^{(s)}) + \sum_{m=1}^{M} (1-\lambda) p(x|s_i^{(s)}, s_{im}^{(d)}) p(s_{im}^{(d)}|s_i^{(s)}), \tag{2}$$

where $s_i^{(s)}$ is ith tied-state in a standard triphone HMM; M is the number of pronunciation variants occurring in dialectal speech for $s_i^{(s)}$; $s_{im}^{(d)}$ is ith state in the mth dialectal monohone HMM; parameter λ is a linear interpolating coefficient between standard and dialectal acoustic models which is usually determined empirically, for example, λ was set to 0.75 in (Tomokiyo, 2001). $p(s_{im}^{(d)}|s_i^{(s)})$ is the probability of the mth pronunciation variant at state level in dialectal speech given a standard state and hence $\sum_{m=1}^{M} p(s_{im}^{(d)}|s_i^{(s)}) \equiv 1$. Afterwards, Eq. (2) can be further simplified and expanded as Eqs. (3) and (4):

$$p'(x|s_i) = \lambda p(x|s_i^{(s)}) + \sum_{m=1}^{M} (1-\lambda) p(x|s_{im}^{(d)}) p(s_{im}^{(d)}|s_i^{(s)}) \tag{3}$$

$$= \sum_{k=1}^{K} \lambda w_{ik}^{(s)} N_{ik}^{(s)}(\cdot) + \sum_{m=1}^{M} \sum_{n=1}^{N} (1-\lambda) \cdot p(s_{im}^{(d)}|s_i^{(s)}) \cdot w_{imn}^{(d)} N_{imn}^{(d)}(\cdot). \tag{4}$$

In Eq. (4), K and N are the numbers of Gaussian mixtures of state $s_i^{(s)}$ and state $s_{im}^{(d)}$, respectively; $w_{ik}^{\prime(s)} = \lambda w_{ik}^{(s)}$ and $w_{imn}^{\prime(d)} = (1-\lambda) \cdot p(s_{im}^{(d)}|s_i^{(s)}) \cdot w_{imn}^{(d)}$ are new mixture weights for standard and dialectal Gaussian mixtures respectively in the merged state of SDPBMM.

From Eq. (4), it can be seen that the new weight for Gaussian mixture from standard speech, $w_{ik}^{\prime(s)}$, is controlled by λ, and the new weight for Gaussians mixture from dialectal speech, $w_{imn}^{\prime(d)}$, is controlled by both the pronunciation variation modeling between standard speech and dialectal speech, $p(s_{im}^{(d)}|s_i^{(s)})$, and λ. Normally, $w_{ik}^{\prime(s)} \gg w_{imn}^{\prime(d)}$, that is, the standard speech has a greater effect on the output PDF in the merged state of SDPBMM. It indicates that, on the one hand, SDPBMM can essentially be regarded as an extension to acoustic model based on standard speech with richer acoustic coverage on dialectal speech. In other words, a dialectal phoneme with some deviation from standard can be covered acoustically. On the other hand, due to the fact that merging takes place between a tied-state of standard triphones and state(s) of dialectal monophones, the dialectal monophones can inherit recognition capability from standard triphones since they are much more precisely and robustly modeled. For example, in Fig. 3, in SDPBMM, the merged state of the Initial *an*, has broader acoustic coverage on the Initial *an* with some deviation resulting from dialect and the *an* sounding similar to *ang* for dialectal speech recognition. Therefore, it is expected that good performance can be achieved for both dialectal speech and standard speech recognition. Also, because SDPBMM keeps the topology of decision tree for standard tied-states unchanged, no modifications to lexicon and decoder are required. Also SDPBMM does not require retraining, which can save time and toil to a great extent.

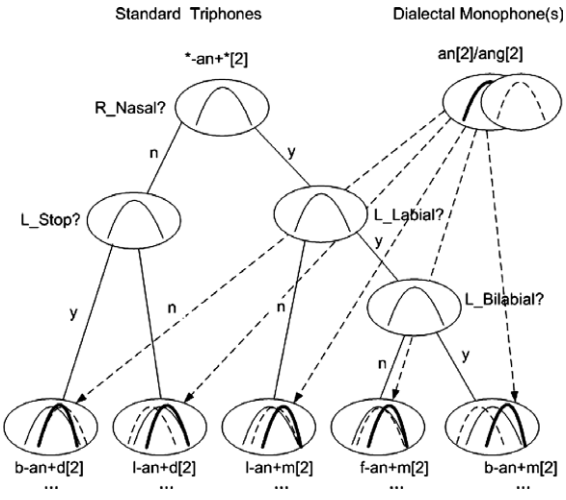

Fig. 3. Unchanged topology after SDPBMM.

4. Pronunciation modeling based on a small amount of dialectal speech

As shown in Eq. (4), $p(s_{im}^{(d)}|s_i^{(s)})$ is the pronunciation variation probability occurring at state level between standard and dialectal HMMs. One of the challenges that

SDPBMM has to face is how to precisely evaluate probability for each pronunciation variant given a small amount of dialectal speech data. Due to data sparseness issue, the pronunciation modeling at the phonetic level is used to estimate the probability at the state level. It is also shown in (Saraclar et al., 2000) that almost no performance difference between phonetic level and state level pronunciation modeling was observed. In pronunciation modeling, knowledge-based and data-driven approaches are widely used (Strik and Cucchiarini, 1999). Typically, the knowledge-based pronunciation modeling is a good choice, especially when no or limited development data is available. However, a drawback is that sometimes there could be a mismatch between the information provided by the phoneticians and the data actually used. Moreover, it is difficult to obtain precise probability for each pronunciation variant which plays an important role in the framework of SDPBMM. In the data-driven pronunciation modeling approach, forced-alignment or phoneme recognition is used to transcribe acoustic signals. Based on the resulting transcription, dynamic programming can be performed to derive mapping rules, build a decision tree, train an artificial neural network, or calculate a phone confusion matrix (Strik and Cucchiarini, 1999). With regard to forced-alignment, a constrained network is necessary and likely pronunciation variants are constrained by the network. One advantage of the forced-alignment is that fewer errors are introduced by the recognizer. However, under-coverage for development data might take place. With regard to phoneme recognition, a phoneme-loop based network is used to recognize speech data (Zheng et al., 2002). Hence, it provides an effective approach to the under-coverage problem. However, many errors are also introduced by the recognizer. It is shown in (Gruhn et al., 2004) that there was a mismatch of 45% at phone level between canonical transcription and surface form transcription via phoneme recognition, and half of the errors were introduced by the recognizer. Under such circumstances, an imprecise probability was likely to be obtained for some pronunciation variants, especially those with fewer observations in dialectal speech. This issue is much severe when only limited development data is available.

Taking these factors into account, we choose the forced-alignment approach to pronunciation modeling in SDPBMM. One prominent issue faced by the forced-alignment approach is how to construct a network which can not only cover likely pronunciation variants but also reduce the errors introduced by a recognizer. In this paper, a distance-based pronunciation modeling (DBPM) approach is proposed as a solution.

Suppose that A is an HMM for a phoneme built on *Putonghua* and A' is the HMM for the same phoneme built on a certain dialectal Chinese. It is assumed that the similarity between Model A and Model A' can be measured by their acoustic distance. The closer they are, the more similar they are; likewise, two less similar models will have a bigger distance acoustically. Thus the distance between the *Putonghua* HMM and its corresponding dialectal HMM can be measured quantitatively. The steps for the DBPM are depicted in Fig. 4.

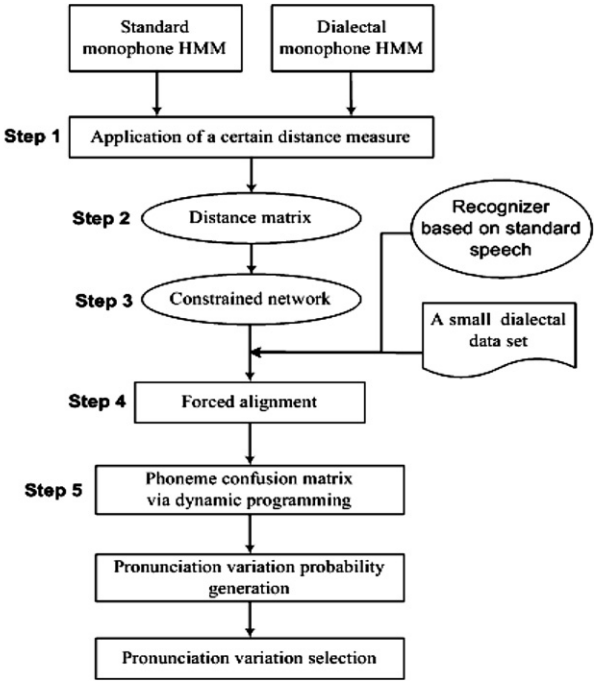

Fig. 4. Flowchart for distance-based pronunciation modeling at phonetics level.

1. Generation of distance matrices for Chinese Initial/Final (IF)[1]
   The distances between one mono-IF HMM from Putonghua and every mono-IF HMM from dialectal speech are calculated. Under the assumption that initials and finals are not mutually confusable, two distance matrices are generated for initial and final sets, respectively. In our study, the Bhattachyaryya distance measure is adopted because it is capable of characterizing the distance more precisely by taking the difference of covariance into account (Huang et al., 2001). Bhattachyaryya distance measure is defined as

$$d(\lambda_1,\lambda_2)=\frac{1}{8}(\mu_1-\mu_2)^T\left(\frac{\Sigma_1+\Sigma_2}{2}\right)^{-1}(\mu_1-\mu_2)+\frac{1}{2}\times\ln\frac{|(\Sigma_1+\Sigma_2)/2|}{|\Sigma_1|^{1/2}|\Sigma_2|^{1/2}}. \quad (5)$$

2. Likely pronunciation variants selection
   The selection is based on the distances obtained in Step 1. In most cases, an IF from *Putonghua* is closest to the exactly same IF from dialectal speech. Usually, the first N dialectal IFs with the shortest distance are selected. N, for example, was set to 4 in our experiment.

[1] Initial/Final in Chinese is similar to phoneme in English, which is commonly used in Chinese speech recognition.

3. Construction of a constrained network.
   The network is constructed based on the selections produced in Step 2. In the resulting network, every IF has N likely pronunciation variants and final pronunciation variants most suitable for the dialectal speech of interest are derived from these N candidates. Afterwards, forced alignment (Lussier, 2003) on a basis of the constrained network using a standard speech recognizer is performed to obtain the most suitable phonetic pronunciation for a certain dialectal Chinese.
4. Phoneme confusion matrix generation.
   The surface form transcription produced by the forced alignment is aligned with canonical transcription produced via syllable-to-phoneme conversion. Consequently, a phoneme confusion matrix is then generated by means of dynamic programming.
5. Selection of pronunciation variants.
   Pronunciation variations between standard speech and dialectal speech can be derived from the resulting confusion matrix. The probability for each pronunciation variation is calculated using Eq. (6)

$$P(d|s) = \frac{N(d,s)}{N(s)} \times 100\%. \tag{6}$$

   To further decrease the confusability, a relative probability is used as a threshold for pruning. Finally, the probabilities for those pronunciation variations surviving the pruning are normalized.

5. Experiments and results

The Mandarin Broadcast News (MBN) database (Hub-4NE) (LDC, 1997), a read style standard Chinese speech corpus, was used to train a baseline system, *i.e. Putonghua* recognizer. It contained about 30 h of high quality wideband speech with detailed Chinese IF transcriptions. The acoustic models of the *Putonghua*-based baseline contained 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. Thirty-nine-dimensional MFCC coefficients with log energy, Δ, and $\Delta\Delta$ were used as features with cepstral mean normalization (Huang et al., 2001). In fact, the same acoustic models were used in (Sproat et al., 2004; Zheng et al., 2005; Li et al., 2006). Also, six zero-initials were added to the standard IF set to help improve the performance and make the modeling process consistent. Another database used here was Wu dialectal Chinese database (WDC) (Li et al., 2003), which contained 100 native Shanghai speakers, 50 males and 50 females. WDC was recorded under a similar condition as that of MBN, therefore, channel mismatch can be minimized. WDC was composed of the read-style speech from medium and strong Shanghai-accented speakers. More details can be found in (Li et al., 2003).

Two data sets, *Train_MBN* and *Test_MBN*, were selected from MBN for training and testing respectively. A test set, *Test_WDC*, was selected from WDC for performance evaluation. These data sets are listed in Table 1 in detail. 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 20.0% on Shanghai-dialectal Chinese speech. With respect to performance of *Putonghua* HMMs, similar results on Shanghai-dialectal Chinese were also achieved in (Sproat et al., 2004; Zheng et al., 2005; Li et al., 2006). Because acoustic modeling was our research focus, no language model was used. Hence, our experiments were performed at Chinese syllable level and SER reduction was used as a measure of system performance. So, a lexicon of 406 toneless Chinese syllables was taken. Besides, HTK 3.2 (Young et al., 2002) was used in these experiments.

Table 1
Data division for the development and test sets

Data set	Database	Detailed information
Train_MBN	MBN	34,500 utterances, approximately 30-h speech
Test_MBN	MBN	1200 utterances, totally 80-min speech
*Dev_WDC*1	WDC	10 speakers, 510 utterances, totally 40-min speech
*Dev_WDC*2	WDC	20 speakers, 1070 utterances, approximately 70-min speech
*Dev_WDC*3	WDC	40 speakers, 2100 utterances, approximately 120-min speech
*Dev_WDC*4	WDC	60 speakers, 3050 utterances, approximately 180-min speech
*Dev_WDC*5	WDC	80 speakers, 3860 utterances, approximately 240-min speech
Test_WDC	WDC	20 speakers, 995 utterances, approximately 60-min speech

It is well known that recognition performance on a certain dialectal speech usually relies on quantity of data used. We intended to use as little dialectal speech data as possible to achieve good performance. Some experiments were performed to determine the appropriate development set for dialectal speech. In this research, another 5 data sets, consisting of 40-min, 70-min, 120-min, 180-min and 240-min dialectal speech, were selected from WDC which are also detailed in Table 1. These five data sets were used to train Shanghai-dialectal Chinese-specific acoustic models from scratch. Here, the same methods as that of *Putonghua* acoustic modeling were adopted. The results evaluated on *Test_WDC* are depicted in Fig. 5. In Fig. 5, it is shown that with the increase of training data from WDC, as expected, better performance in dialectal Chinese speech recognition was achieved. In the best case, the training set, *Dev_WDC*5, consisting of all speech data from WDC except *Test_WDC*, achieved an SER of 38.3%. In addition, a trend observed through Fig. 5 is that the SER achieved by retraining on the basis of training sets of more than one hour was lower than the baseline. In other words, if more than one hour of dialectal speech data is given, acoustic models built by retraining using only Shanghai-dialectal speech data outperformed those built using more than 30 h *Putonghua* speech data. Simply put, one-hour data can be a

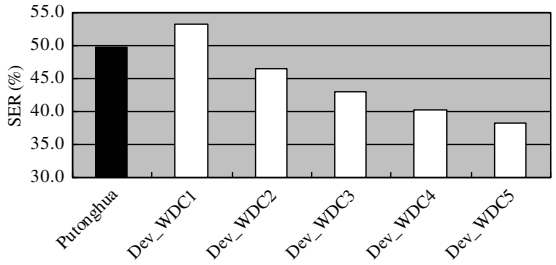

Fig. 5. Comparison between dialect-specific retraining and *Putonghua* baseline evaluated on *Test_WDC*.

threshold below which the data set can be regarded as *a small data set*. Thus, we focus on how to make full use of a dialectal speech data set less than one hour for dialectal Chinese speech recognition. In later experiments, 40-min data set, *Dev_WDC*1, was used as default development set for SDPBMM-based acoustic modeling.

The data set, *Dev_WDC*1, was used to build 65 context-independent dialectal mono-IF HMMs for SDPBMM, each mono-IF HMM was of the exactly same topology as that of standard tri-IF HMM except that there were six Gaussian mixtures per state.

5.1. Distance-based pronunciation modeling

In distance-based pronunciation modeling, the standard mono-IF HMMs were built based on *Train_MBN*, which were the intermediate results of context-dependent acoustic modeling, while the dialectal mono-IF HMMs were built based on *Dev_WDC*1. Within the constrained network, the 4 most likely candidates for each canonical pronunciation were included in terms of their distances. The *Putonghua* acoustic model in combination with the constrained network performed the forced-alignment on *Dev_WDC*1 so as to obtain surface form transcription for dialectal Chinese. A relative probability of 15%, was used as a threshold for pruning. By following the steps described in Fig. 4, final pronunciation variants were obtained, some of which with multiple pronunciation variants are listed in Table 2.

In fact, the resulting pronunciation variants listed in Table 2 are quite consistent with the phonetic knowledge about Shanghai-dialectal Chinese (Li et al., 2006; Li and Wang, 2003). For example, in Shanghai dialect, there exist no retroflex Initials, [*zh, ch, sh*]. Instead they are pronounced as [*z, c, s*] in Shanghai-dialectal Chinese. The pronunciation variations, /*zh* → *z*/, /*ch* → *c*/, and /*sh* → *s*/ are of high pronunciation variation probability in Shanghai-dialectal speech; but not vice versa. Another example is that [*en*] and [*eng*] are mutually confusable to a higher degree in Shanghai-dialectal Chinese, so almost equal pronunciation variation probability is achieved for /*eng* → *en*/ and /*en* → *eng*/.

5.2. Evaluations on SDPBMM – related acoustic models

In SDPBMM, the development set, *Dev_WDC*1, had been used to build the dialectal mono-IF HMMs in Section 5.1; the linear coefficient λ in Eq. (3) was determined experimentally and was set to 0.72. In addition, the pronunciation modeling weight, $p(s_{im}^{(d)}|s_i^{(s)})$, as listed in Table 2, was also incorporated into SDPBMM in accordance with Eq. (4). To make it clear, the overall procedure for evaluating the effectiveness of SDPBMM is comprehensively depicted in Fig. 6. A series of experiments were designed and performed according to it.

Table 2
Some initials/finals with multiple pronunciation variants obtained via distance-based pronunciation modeling

Canonical	Probability	Surface form
c	0.736	c
	0.264	ch
en	0.572	en
	0.428	eng
ia	0.731	ia
	0.269	iang
ii	0.764	ii
	0.236	iii
in	0.432	in
	0.568	ing
s	0.729	s
	0.271	sh
uan	0.753	uan
	0.247	an
ve	0.750	ve
	0.250	ie
ch	0.546	c
	0.454	ch
eng	0.564	eng
	0.436	en
ie	0.797	ie
	0.203	ve
iii	0.571	ii
	0.429	iii
ing	0.333	in
	0.667	ing
sh	0.477	sh
	0.523	s
zh	0.624	z
	0.376	zh
r	0.716	r
	0.284	l

The basic components for *Putonghua* acoustic model AM0 and SDPBMM-based acoustic model AM1 are listed in Table 3. Due to the fact that the model merging was performed at state level, only the number of Gaussian mixtures of each state was increased. Compared with AM0, the overall number of Gaussian mixtures was increased by approximately 53% in AM1.

Two acoustic models, AM0 and AM1, were evaluated by *Putonghua* set *Test_MBN*, and Shanghai-dialectal Chinese set *Test_WDC*, respectively. The results are listed in Table 4.

It shows that SDPBMM can give an SER reduced by an absolute 7.1% compared with AM0 when *Test_WDC* was used. It can also achieve a slightly higher SER of only 0.9% on *Test_MBN* than AM0. That is to say, SDPBMM can achieve a significant improvement in dialectal speech

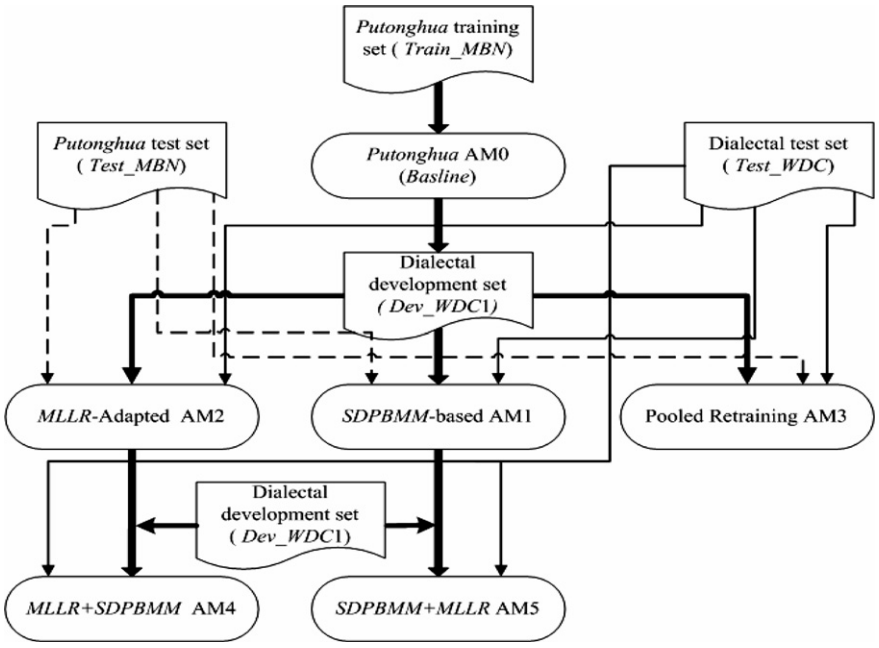

Fig. 6. Overall procedure for SDPBMM-related evaluation.

Table 3
Basic components for *Putonghua* (AM0) and SDPBMM (AM1) acoustic models

	Putonghua	SDPBMM
States	3230	3230
Gaussian mixtures	45,220	68,866
Tri-IF HMMs	7411	7411

Table 4
Comparisons of *Putonghua* (AM0) and SDPBMM (AM1) on *Test_MBN* and *Test_WDC*

Acoustic model	SER	
	Test_MBN (%)	*Test_WDC* (%)
AM0	30.5	49.8
AM1	31.4	42.7

recognition without significant degradation in standard speech recognition.

In AM1, it is naturally assumed that improvement in dialectal speech recognition may result from increase in Gaussian mixtures in the merged states. Compared to AM0, where 14 Gaussian mixtures per state are contained, on average, there were 21.3 mixtures per state in AM1. To make a fair comparison, another *Putonghua* acoustic model with 22 Gaussian mixtures per state was built on *Train_MBN,* which had approximately equal parameter scale as that of AM1. The SER on *Test_WDC* was decreased from 49.8% to 49.0%. However, compared to AM1, a gap of an absolute 6.3% still exists. These results show that significant improvement cannot be achieved by simply increasing the parameter scale for dialectal speech recognition.

5.3. Comparison of SDPBMM with pooled retraining

It was reported in (Wang et al., 2003) that significant recognition improvement on non-native speech was achieved by'pooled' retraining. To make a fair comparison of SDPBMM and the pooled retraining, exactly the same amount of dialectal data, *Dev_WDC*1, pooling with *Train_MBN*, was used for the retraining. Same modeling procedure as AM0 was adopted and the resulting acoustic model was named AM3, as indicated in Fig. 6. The results of AM3 on *Test_MBN* and *Test_WDC* are listed in Table 5.

Through the pooled retraining, the SER of AM3 on *Test_WDC* was decreased from 49.8% to 48.0% whilst the SER on *Test_MBN* was increased from 30.5% to 31.3%. Compared with AM1, a comparable performance on *Putonghua* was achieved but an absolute gap of 5.3% in SER on dialectal Chinese still existed. It shows that retraining by pooling a small amount of dialectal speech cannot achieve as significant improvement as SDPBMM on dialectal Chinese. In pooled retraining, some triphone HMMs were built only on standard speech due to severe data sparseness of dialectal speech. No accented speech training samples were available for these HMMs into which no information from dialectal speech could be incorporated. Instead, in SDPBMM, Gaussian mixtures from

Table 5
Results for pooled retraining (AM3) on *Test_MBN* and *Test_WDC*

Acoustic model	SER	
	Test_MBN	*Test_WDC*
AM3	31.3%	48.0%

dialectal speech were explicitly merged into the states from standard speech. Much information from dialectal speech was transformed into acoustic models of standard speech.

5.4. Comparison of SDPBMM with adaptation method

It was reported in (Wang et al., 2003) that MLLR is much beneficial when only a small amount of data is available, so MLLR adaptation was performed based on *Dev_WDC*1 and AM0. In MLLR adaptation, all the standard tri-IFs were classified into 65 classes, and only mean update was performed in transformation matrix. The resulting acoustic model was denoted as AM2 in Fig. 6. As a result, an SER of 44.1% was achieved on *Test_WDC* which was still higher than the SER of 42.7% by AM1 with exactly same data set. These results are listed in Table 6. It shows that compared with MLLR, SDPBMM can achieve better performance (with an absolute SER reduction of 1.4%) on dialectal speech. Also, as expected, the adapted acoustic model usually worsens the performance when evaluated on *Putonghua*, the SER on *Test_MBN* was increased by 13.3%. It shows that only applying adaptation over standard acoustic model cannot achieve good performance on dialectal speech and standard speech simultaneously.

5.5. Integration of SDPBMM with adaptation

Naturally, a question arises whether SDPBMM can give good results when being integrated with some existing techniques. To make it clear, SDPBMM was integrated with adaptation.

It is assumed that SDPBMM primarily concentrates on addressing the issues of the phonetic mismatch between dialectal speech and standard speech. Because the adaptation has been proven as a good solution to channel mismatch, it is expected that in combination with a certain adaptation method, SDPBMM can further improve the performance on dialectal speech. To verify the assumption, *Dev_WDC*1, was used for adaptation again. As a result, two new acoustic models SDPBMM + MLLR and MLLR+SDPBMM were built from AM5 and AM4, respectively, as shown in Fig. 6. They were performed in the same order as mentioned here. For example, in AM5, the SDPBMM was performed on AM0 followed by the MLLR adaptation using *Dev_WDC*1. The results are listed in Table 7. From the table, it can be seen that through the integration of SDPBMM and MLLR adaptation, further SER reductions of 0.5% and 1.4% on dialectal Chinese speech were achieved by AM4 and AM5, respectively. It is shown that SDPBMM and adaptation methods can act as complementary procedures for each other. Nevertheless, AM5 outperformed AM4 by 0.9%, so SDPBMM is more appropriate to be a front-end component for adaptation in ASR system.

Table 6
Results for MLLR adaptation (AM2) on *Test_MBN* and *Test_WDC*

Acoustic model	SER	
	Test_MBN	*Test_WDC*
AM2	43.8%	44.1%

Table 7
Results for MLLR + SDPBMM (AM4) and SDPBMM + MLLR (AM5) on *Test_MBN* and *Test_WDC*

Acoustic model	SER (%)
AM4	42.2
AM5	41.3

6. Conclusions and future work

In our paper, SDPBMM, a novel, simple and effective acoustic modeling method for dialectal Chinese speech recognition, is proposed. To obtain new mixture weights for DPBMM constrained by pronunciation modeling, distance-based pronunciation modeling is proposed especially based on a small amount of dialectal speech data. Through a series of experiments, it is concluded that SDPBMM possesses the following advantages: (1) it is simple but practical for acoustic modeling when only a small amount dialectal speech data is available; (2) it can achieve a significant performance improvement on dialectal Chinese; (3) it can show good performance for both standard and dialectal speech recognition; (4) it can achieve a better performance than adaptation, especially when given a small amount of dialectal speech data; (5) it is complimentary to adaptation, that is to say, SDPBMM together with adaptation can improve performance for dialectal speech recognition. In a word, SDPBMM is one of the most effective acoustic modeling methods for read-style dialectal Chinese speech recognition when small amount of data is available.

In this paper, the experiments were conducted on Shanghai-dialectal Chinese, and yet no dialect-specific knowledge was incorporated in acoustic modeling, thus, the proposed methods can be easily generalized to other dialectal Chinese speech recognition.

One apparent issue in SDPBMM is that when performing model merging, the number of Gaussian mixtures in a merged state of SDPBMM is increased significantly. For example, in our study, the number of Gaussian mixtures is increased by 53% while time cost for decoding procedure is increased by almost 60%. We refer to this issue as Gaussian mixture expansion problem. How to select the Gaussian mixtures with strong discriminative ability to involve in merging process is a potential debate for future.

Another issue is that all the experiments were performed based on read speech. In our next step research on spontaneous/conversational speech will be carried out. Because more complicated pronunciation variability exists in spontaneous dialectal Chinese, more robust *Putonghua*

recognizer would be required to work in integration with SDPBMM to reap its benefits.

References

Angkititrakul, P., Hansen, J.-H.-L., 2006. Advances in phone-based modeling for automatic accent classification. IEEE Trans. on Audio, Speech Language Processing 14 (2), 634–646.

Chen, T., Huang, C., Chang, E. and Wang, J.-C., 2001. Automatic accent identification using Gaussian mixture models. In: Proc. ASRU.

Diakoloukas,V., Digalakis, V., Neumeyer, L., Kaja, J., 1997. Development of dialect-specific speech recognizers using adaptation methods. In: Proc. ICASSP.

Fung, P., Liu, Y., 2005. Effects and modeling of phonetic and acoustic confusions in accented speech. J. Acoust. Soc. Amer. 118 (5), 3279–3293.

Gao, J.-F., Goodman, J., Li, M.-J., Lee, K.-F., 2002. Toward a unified approach to statistical language modeling for Chinese. ACM Trans. Asian Language Inform. Process. 1 (1), 3–33.

Goronzy, S., Kompe, R., Rapp, S., 2004. Generating non-native pronunciation variants for lexicon adaptation. Speech Comm. 42 (1), 109–123.

Gruhn, R., Markov, K., Nakamura, S., 2004. A statistical lexicon for non-native speech recognition. In: Proc. ICSLP.

He, X.-D., Zhao, Y.-X., 2003. Fast model selection based speaker adaptation for non-native speech. IEEE Trans. Speech Audio Process. 11 (4), 298–307.

Huang, X.-D., Acero, A., Hon, S.-W., 2001. Spoken Language Processing. Prentice Hall.

Huang, C., Chen, T., Chang, E., 2004. Accent issue in large vocabulary continuous speech recognition. Internat. J. Speech Technol. 7, 141–153.

Huang, R.-Q., Hansen, J.-H.-L., 2005. Dialect/accent classification via boosted word modeling, In: Proc. ICASSP, pp. 585–588.

Hwang, M.-Y., Huang, X.-D., Alleva, F.-A., 1996. Predicting unseen triphones with senones. IEEE Trans. Speech Audio Process. 4 (6), 412–419.

LDC, 1997. <http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2007T19>.

Li, A.-J., Wang, X., 2003. A contrastive investigation of standard mandarin and accented mandarin. In: Proc. EuroSpeech.

Li, J., Zheng, F., Xiong, Z.-Y., Wu, W.-H., 2003. Construction of large-scale Shanghai Putonghua speech corpus for Chinese speech recognition. In: Proc. Oriental-COCOSDA, pp. 62–69.

Li, J., Zheng, T.-F., Byrne, W., Jurafsky, D., 2006. A dialectal Chinese speech recognition framework. J. Comput. Sci. Technol. 21 (1), 106–115.

Liu, Y., Fung, P., 2004. Pronunciation modeling for spontaneous mandarin speech recognition. Internat. J. Speech Technol. 7, 155–172.

Livescu, K., 1999. Analysis and modeling of non-native speech for automatic speech recognition. Master Thesis, Massachusetts Institute of Technology.

Lussier, E.-F., 2003. A tutorial on pronunciation modeling for large vocabulary speech recognition. Lect. Notes Comput. Sci. 2705, 38–77.

Myrvoll, T.-A., 2003. Adaptation techniques in automatic speech recognition. Telektronikk 2, 59–69.

Oh, Y.-R., Yoon, J.-S., Kim, H.-K., 2007. Acoustic model adaptation based on pronunciation variability analysis for non-native speech recognition. Speech Comm. 49, 59–70.

Saraclar, M., Nock, H., Khudanpur, S., 2000. Pronunciation modeling by sharing gaussian densities across phonetic models. Comput. Speech Language 14, 137–160.

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., 2004. Dialectal Chinese speech recognition: final technical report. <http://www.clsp.jhu.edu/ws2004/>.

Strik, H., Cucchiarini, C., 1999. Modeling pronunciation variation for ASR: a survey of the literature. Speech Comm. 29, 225–246.

Tjalve, M., Huckvale, M., 2005. Pronunciation variation modeling using accent features. In: Proc. EuroSpeech.

Tomokiyo, L.-M., 2001. Recognizing non-native speech: characterizing and adapting to non-native usage in LVCSR. Ph.D. Thesis, Carnegie Mellon University, USA.

Wang, Z.-R., Schultz, T., Waibel, A., 2003. Comparison of acoustic model adaptation techniques on non-native speech. In: Proc. ICASSP, pp. 540–543.

Xu, X.-T., Yan, Y.-H., 2004. Speaker adaptation using constrained transformation. IEEE Trans. Speech Audio Process. 12 (2), 168–174.

Young, S., Evermann,G., Hain,T., et al., 2002. The HTK Book (for HTK Version 3.2.1). Cambridge University. <http://htk.eng.cam.ac.uk/>.

Zheng, F., Song, Z.-J., Fung, P., Byrne, W., 2002. Mandarin pronunciation modeling based on CASS corpus. J. Comput. Sci. Technol. 17 (3), 249–263.

Zheng,Y.-L., Sproat, R., Gu, L., et al., 2005. Accent detection and speech recognition for Shanghai-accented mandarin. In: Proc. EuroSpeech.