Using a Small Development Set to Build a Robust Dialectal Chinese Speech Recognizer

Linquan Liu[1], *Thomas Fang Zheng*[1], *Makoto Akabane*[2], *Ruxin Chen*[3], *Wenhu Wu*[1]

[1]Center for Speech Technology, Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing, 100084, China
[2]Sony Computer Entertainment Inc., Tokyo, 107-0062, Japan
[3]Sony Computer Entertainment America, Foster City, CA, 94404, USA

liulq@cst.cs.tsinghua.edu.cn, fzheng@tsinghua.edu.cn, akabane@rd.scei.sony.co.jp, Ruxin_Chen@PlayStation.Sony.com, wuwh@tsinghua.edu.cn

Abstract

To make full use of a small development data set to build a robust dialectal Chinese speech recognizer from a standard Chinese speech recognizer (based on Chinese Initial/Final, IF), a novel, simple but effective acoustic modeling method, named state-dependent phoneme-based model merging (SDPBMM), is proposed and evaluated, where a shared-state of standard tri-IF is merged with a state of dialectal mono-IF in terms of pronunciation variation modeling. Specifically, in order to deal with phonetic-level pronunciation variations in SDPBMM, distance-based pronunciation modeling is proposed based on a small dialectal Chinese data set. With a 40-minute Shanghai-dialectal Chinese data set, SDPBMM can achieve a significant syllable error rate (SER) reduction of 14.3% for dialectal Chinese with almost no performance degradation for standard Chinese. Experimentally, SDPBMM can also outperform the maximum likelihood linear regression (MLLR) adaptation and the pooled retraining methods with relative SER reductions by 2.8% and 10.6%, respectively. If SDPBMM is combined with the MLLR adaptation, another relative SER reduction of 3.3% can be further achieved.

Index Terms: dialectal Chinese, speech recognition, accented speech, pronunciation modeling, acoustic modeling

1. Introduction

Accent is one of the challenges in current automatic speech recognition (ASR) systems. Dialectal speech is the speech with similar or identical accent which is of some regional characteristics. Nowadays, as for accented/dialectal speech recognition, there are several aspects on which a great deal of research has been carried out. 1) Pronunciation modeling. The pronunciation lexicon is one of the principal targets on which most work has been focused [1]. Specifically, some state-level pronunciation modeling efforts have also been made [2]. 2) Retraining [3]. Some retraining mechanisms have been proposed where standard speech and dialectal/accented speech are pooled together to build a robust ASR system. 3) Adaptation [4]. The adaptation technique is an extremely effective way to improve system performance on accented speech recognition. 4) Decoder tuning [5]. Modifications are made to the decoder to better characterize accented/dialectal speech. 5) Accent classification [6]. It is usually employed as a front-end for ASR systems. In practice, the aforementioned approaches are often combined to build a robust ASR system.

In China, *Putonghua* (standard Chinese) is an official language through which Chinese people from different regions can be mutually understood. *Putonghua* spoken by most Chinese people is usually influenced by their native dialect more or less. In this paper, *Putonghua* influenced by a certain Chinese dialect is referred to as dialectal Chinese. In general, it is impractical to collect a large amount of data to build a recognizer for each dialectal Chinese due to its diversity, therefore one of our motivations here is to build a robust recognizer for a dialectal Chinese based on a handy *Putonghua* recognizer along with a small dialectal Chinese data set (less than one hour). Another important motivation here is to make the built recognizer work well for both dialectal and standard speech recognition simultaneously.

To address these issues, we propose an acoustic modeling approach named state-dependent phoneme-based model merging (SDPBMM) where based on a certain Chinese Initial or Final (IF), Gaussian mixtures at state level from a context-dependent *Putonghua* tri-IF HMM and its IF-related context-independent dialectal mono-IF HMM(s) are merged according to pronunciation variation modeling between them. To a great extent, the newly-merged HMM can represent both dialectal and standard speech characteristics acoustically. Acting as a merging criterion, the state-level pronunciation modeling plays an important role in SDPBMM. Accordingly, how to capture the pronunciation variants and precisely evaluate their pronunciation variation probability based on a small amount of data is an issue due to the fact of data sparseness. Accordingly, a distance-based pronunciation modeling method based on a small data set is proposed. In addition, as a side effect of SDPBMM, the number of Gaussian mixtures within the merged states will be increased definitely, which is referred to as the Gaussian mixture expansion problem. To downsize the scale of Gaussian mixtures while without causing any degradation on dialectal Chinese speech, the states that need merging must be differentiated from those that do not need merging in SDPBMM. Therefore, a distance measure, named *pseudo-divergence based distance measure* (PDBDM), is proposed to solve it. .

In this paper, only 40-minute Shanghai-dialectal speech data is adopted to build a cost-effective acoustic model for the Shanghai-dialectal Chinese from a *Putonghua* recognizer via the proposed methods.

2. State-dependent phoneme-based model merging

To take both *Putonghua* and dialectal Chinese into account at state level, the SDPBMM is performed. Most of the state-of-the-art Chinese ASR systems tend to use context-dependent

tri-IF (similar to triphone in English) HMMs and a decision tree based state sharing method to build robust acoustic models [7]. Keeping the topology of a decision tree unchanged, we attempt to merge Gaussian mixtures from context-independent HMM(s) for dialectal Chinese into its IF-related context-dependent HMM for *Putonghua* at state level within a decision tree, this is the so-called SDPBMM. The basic idea of SDPBMM can be illustrated in Figure 1. In the left part of Figure 1, taking a Chinese Final *an* as an example, the 2nd states from *an*-centered tri-IFs are presented by a decision tree. To accomplish the merging process, the 2nd states from the dialectal mono-IF *an* and one of its pronunciation variants *ang* are merged with the leaf nodes of *an*-centered decision tree, *i.e.* the shared-states. In that case, whether a pronunciation variant, *ang* of *an*, should be involved in the merging or not is determined by the pronunciation modeling which will be introduced in Section 3. The merging process is depicted in the right part of Figure 1. As a result, a merged shared-state consists of multiple Gaussian mixtures from both a state of *Putonghua* tri-IF HMM and its corresponding state of dialectal mono-IF HMM, as denoted by thin black curves and thick black curves in Figure 1, respectively.

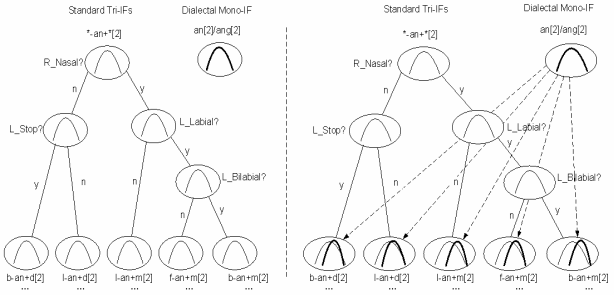

Figure 1: *The basic idea of SDPBMM.*

Theoretically, SDPBMM is formulated as follows. Let x and s_i be an input vector and the i-th state in a HMM, respectively, then a probability density function (*pdf*) for a continuous HMM $p(x|s_i)$ is formulated by Equation 1. For simplification, $N_{ik}(\cdot)$ will be used to denote $N(x; \mu_{ik}; \Sigma_{ik})$ for the i-th state hereinafter.

$$p\left(x\left|s_i\right.\right)=\sum_{k=1}^{K} w_{ik} N\left(x;\mu_{ik};\Sigma_{ik}\right) \quad (1)$$

Let $p'(x|s_i)$ be the modified *pdf* for a merged shared-state after applying SDPBMM, which can be represented as

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

where $s_i^{(s)}$ is the i-th shared-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 the i-th state in the m-th dialectal monohone HMM, and parameter λ is a interpolating coefficient between standard and dialectal acoustic models. In fact, $p\left(s_{im}^{(d)}\left|s_i^{(s)}\right.\right)$ is the probability of the m-th pronunciation variant at state level in dialectal speech given a standard state. Afterwards, Equation 2 can be further simplified and expanded as

$$p'(x\left|s_i\right.)=\sum_{k=1}^{K}\lambda w_{ik}^{(s)} N_{ik}^{(s)}(\cdot)+\sum_{m=1}^{M}\sum_{n=1}^{N}(1-\lambda)\cdot P(s_{im}^{(d)}\left|s_i^{(s)}\right.)\cdot w_{imn}^{(d)} N_{imn}^{(d)}(\cdot) \quad (3)$$

where K and N are the numbers of Gaussian mixtures of states $s_i^{(s)}$ and $s_{im}^{(d)}$, respectively. $w'^{(s)}_{ik}=\lambda w_{ik}^{(s)}$ and $w'^{(d)}_{imn}=(1-\lambda)\cdot p\left(s_{im}^{(d)}\left|s_i^{(s)}\right.\right)\cdot w_{imn}^{(d)}$ are new mixture weights for standard and dialectal Gaussian mixtures respectively in the merged state of SDPBMM. $w_k^{(s)}$ is controlled by both the original weight and λ; likewise, the new weight from dialectal speech, $w'^{(d)}_{imn}$, is controlled by the original weight, pronunciation variation probability $p\left(s_{im}^{(d)}\left|s_i^{(s)}\right.\right)$ and λ. Normally, $w'^{(s)}_{ik}\gg w'^{(d)}_{imn}$, that is to say, a standard state has a greater effect on the output *pdf* in the merged state. It indicates that SDPBMM can be essentially regarded as an extension of standard speech based acoustic model into one with richer acoustic coverage on dialectal speech, and therefore it can be expected to achieve good recognition performance for both dialectal speech and standard speech.

3. Pronunciation modeling based on a small amount of dialectal speech data

In Equations 2 and 3, $p\left(s_{im}^{(d)}\left|s_i^{(s)}\right.\right)$ is the pronunciation modeling at state level. One of the challenges that SDPBMM has to face is how to more precisely estimate the probability for each pronunciation variant given a small amount of dialectal speech data. Because of the data sparseness issue, pronunciation modeling at phonetic level is used to estimate the probability at state level. Generally speaking, knowledge could be more useful for pronunciation modeling especially when no or only limited development data is available [8]. However here for SDPBMM, it is hard to obtain a precise probability for each pronunciation variant. Another pronunciation modeling approach is data-driven where forced-alignment or phoneme recognition is widely used [8]. As for the forced-alignment, a constrained recognition network is necessary and thereby some likely pronunciation variants are under its constraints. One advantage of the forced-alignment method is that fewer errors could be introduced by a recognizer; nevertheless, sometimes under-coverage for development data might take place [9]. For the phoneme recognition, a phoneme-loop network is used. Though it can deal well with the under-coverage issue, more errors are usually introduced by the recognizer [9]. This problem is even more severe when only a limited amount of development data is available.

Taking these factors into account, we choose the forced-alignment pronunciation modeling approach in SDPBMM. An issue here is how to construct a network that can not only cover some likely pronunciation variants but also reduce errors introduced by the recognizer. Consequently, a distance-based pronunciation modeling approach is proposed as a solution.

It is assumed that the similarity between a *Putonghua* HMM and a dialectal HMM can be measured by their acoustic distance. The closer they are, the more similar they are; or equally two less similar HMMs will have a bigger acoustic distance. Some major steps for the distance-based pronunciation modeling are described as follows.

1. Generation of distance matrices. The distance between any one mono-IF HMM from *Putonghua* and any mono-IF HMM from dialectal speech is calculated based on the Bhattachyaryya distance measure. Two distance matrices are generated for Chinese Initial and Final sets, respectively

under the assumption that an Initial and a Final are not mutually confusable in reality. The adoption of Bhattachyaryya distance measure lies in the assumption that it can evaluate the distance between dialectal and standard Chinese mono-IF symmetrically and precisely [2]. The Bhattachyaryya distance measure is defined as

$$d(\lambda_1,\lambda_2)=\frac{1}{8}(\mu_1-\mu_2)^T\left(\frac{\Sigma_1+\Sigma_2}{2}\right)^{-1}(\mu_1-\mu_2)+\frac{1}{2}\ln\frac{\left|(\Sigma_1+\Sigma_2)/2\right|}{|\Sigma_1|^{1/2}|\Sigma_2|^{1/2}} \quad (4)$$

2. Construction of the constrained network. Usually, the first *N* dialectal IFs with shortest distances are selected to construct the constrained network. In most cases, an IF from *Putonghua* is often closest to the same IF from dialectal speech. Accordingly, the final pronunciation variants most suitable for the dialectal speech of interest will be derived from these *N* candidates in terms of the resulted network.

After the constrained network required by the forced-alignment has been constructed according to acoustic distance, the conventional forced-alignment on the basis of constrained network in combination with a *Putonghua* recognizer will be performed to complete the phonetic-level pronunciation modeling [8].

4. Gaussian mixture expansion problem

Though SDPBMM is effective for dialectal Chinese speech recognition, it will also introduce a Gaussian mixture expansion problem, which will certainly lead to much time-consumption during decoding. To solve it, the differentiation of some states that need merging from those that need no merging is necessary in SDPBMM. Therefore, a distance measure, named *pseudo-divergence based distance measure* (PDBDM) [13] is proposed. Practically, a state from a dialectal mono-IF HMM and its corresponding state on a basis of the same IF from a standard tri-IF HMM form a pair for the calculation of distance. The distances of all pairs are computed using PDBDM. 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 under the assumption that in doing so greater acoustical coverage for dialectal Chinese can be achieved. In addition, the threshold in PDBDM is usually defined experimentally.

5. Experiments and results

5.1. Experimental setup

The Mandarin Broadcast News (MBN) database (Hub4NE) [6] was used to train a *Putonghua* recognizer, taken as the baseline in this paper. The acoustic model of the baseline was shared-state cross-word standard tri-IF HMMs using HTK3.2 [10]. Each tri-IF was modeled using a left-to-right 3-state continuous HMM, with 14 Gaussian mixtures per state. Features were 39-dimensional standard MFCC coefficients. The Wu dialectal Chinese database (WDC) [4] with speech recorded under a similar condition to MBN from 50 male and 50 female Shanghai native speakers was also used in the experiments. The WDC was composed of read-style speech from medium and strong Shanghai-accented speakers.

Two data sets, *Train_MBN* (30-hour speech) and *Test_MBN* (1.2-hour speech), were selected from MBN for training and testing, respectively. A 40-minutue dialectal Chinese development set, *Dev_WDC*, and a 1.0-hour test set, *Test_WDC*, were selected from WDC. The *Dev_WDC* was used to build 65 context-independent dialectal mono-IF HMMs for SDPBMM, the mono-IF HMMs was of the same topology as that of *Putonghua* tri-IFs except that there were 6 Gaussian mixtures per state. No language model was used because this paper focuses on acoustic modeling only. Experimental results were evaluated at Chinese syllable level and the Chinese SER reduction was used as a measure for system improvement. A recognition lexicon of 406 toneless Chinese syllables was adopted. The overall procedure is depicted in Figure 2. A series of experiments were designed and conducted according to this figure.

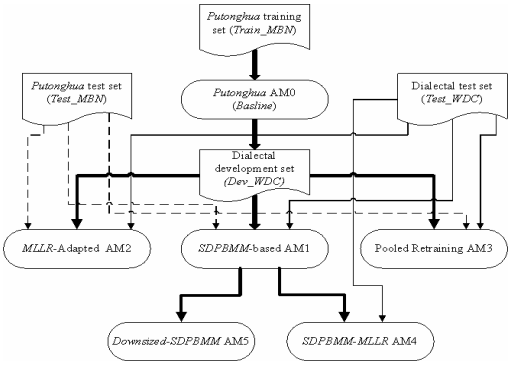

Figure 2: *Procedures for SDPBMM-related evaluations*

In the distance-based pronunciation modeling, the standard context-independent mono-IF HMMs were built using *MBN_Train* while the dialectal mono-IF HMMs using *Dev_WDC*. The constrained network contained 4 most likely candidates with shortest distances for each canonical pronunciation. The *Putonghua* recognizer in combination with the constrained network was used to perform the forced-alignment on *Dev_WDC* to obtain surface form transcriptions for dialectal Chinese. In our study, a relative probability of 15% was used as a threshold for pruning. By following the steps described in Section 3 and [8], finally there were 18 IFs each with two pronunciation variants while others each with only one. In fact, the resulted pronunciation variants were quite consistent with the phonetic knowledge about Shanghai-dialectal Chinese [11].

5.2. Evaluations on acoustic models

The corresponding results evaluated on *Test_MBN* and *Test_WDC* for those acoustic models listed in Figure 2 are presented in Figure 3. In Figure 3, Column *Baseline* is denoted by AM0. Initially, AM0 achieved the SERs of 30.5% and 49.8% on *Test_MBN* and *Test_WDC*, respectively. Column *SDPBMM* is represented by AM1. For AM1, the dialectal 65 mono-IF HMMs was built based on *Dev_WDC*. In SDPBMM, the interpolating coefficient λ was determined experimentally and was optimally set to 0.72 in this paper. In addition, the pronunciation modeling weight, $p\left(s_{im}^{(d)}\middle|s_i^{(s)}\right)$, was also integrated into SDPBMM with the pronunciation variation probabilities obtained via the distance-based pronunciation modeling. Considering that adaptation is one of the most effective methods for accented speech recognition and that MLLR is much beneficial when only a small amount of data available [12], we adopted MLLR for adaptation method based on *Dev_WDC* and AM0. The resulted acoustic model is denoted by AM2 in Figure 2 and represented by Column *MLLR* in Figure 3 accordingly. It is reported in [12] that significant improvement on non-native speech was

achieved by the pooled retraining. For comparison, *Train_MBN* pooled with *Dev_WDC* was used to perform retraining which is denoted by AM3 and Column *Pooled Retraining* in Figures 2 and 3, respectively.

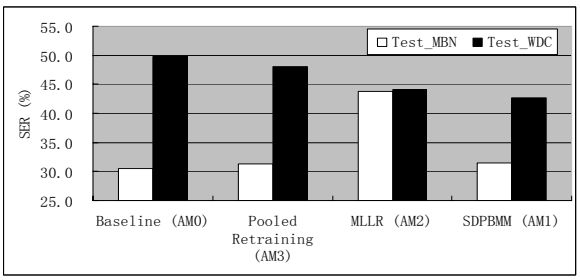

Figure 3: *Evaluations on acoustic models built on Dev_WDC.*

From Figure 3, it can be seen that: 1) SDPBMM could achieve the best performance on *Test_WDC* with a significant relative SER reduction of 14.3% compared with the baseline. It outperformed the MLLR adaptation and the pooled retraining with relative SER reductions of 2.8% and 10.6%, respectively; 2) For *Test_MBN*, as expected, the best performance was achieved by the baseline. However SDPBMM and the pooled retraining also showed good performance on *Putonghua* with absolute SER increases of 0.9% and 0.8% only, respectively. Adaptation method leaded to severe performance degradation on *Putonghua* with an absolute SER increase of 13.3%. It is shown statistically that SDPBMM can achieve good performance on dialectal Chinese without decreasing the performance on *Putonghua*.

5.3. Integration of SDPBMM with adaptation

SDPBMM is primarily proposed to concentrate on addressing the issues of phonetic mismatch between dialectal speech and *Putonghua*. Because the adaptation has been an effective solution to channel mismatch, it is expected that SDPBMM in combination with a certain adaptation method can potentially further improve the recognition performance for dialectal speech. To verify the assumption, *Dev_WDC*, was further used for adaptation and an acoustic model named AM4 was obtained via MLLR adaptation from AM1. An SER of 41.3% on *Test_WDC* was achieved with another relative SER reduction of 3.3% compared with AM1. It is shown that SDPBMM can act as a complement procedure for adaptation, to some extent it can be a front-end component for adaptation.

5.4. Downsize Gaussian mixtures in SDPBMM

To deal specifically with the Gaussian mixtures expansion problem where the overall number of Gaussian mixtures in AM1 was, for example, about 53% more than that in AM0, the PDBDM was adopted and a threshold is set to 0.7. The resulted acoustic model corresponds to *AM5* in Figure 2. The results for downsized-SDPBMM are listed in Table 1.

Table 1. *Evaluations for downsized-SDPBMM AM5*

test set / model	SER	
	Test_MBN	Test_WDC
AM5	31.4%	43.0%

Experimentally, the scale of Gaussian mixtures was decreased by an optimal percentage of 30% with a slight SER increase of 0.3% absolutely on *Test_WDC*; meanwhile, there was no degradation on *Test_MBN* .

6. Conclusion

In the paper, to make full use of a small development data set in dialectal Chinese speech recognition, SDPBMM is proposed. To obtain new mixture weights for SDPBMM constrained by pronunciation modeling, the distance-based pronunciation modeling is proposed specifically based on a small amount of dialectal speech data. To deal with Gaussian mixture expansion problem, PDBDM is used as a method to downsize the scale of Gaussian mixtures in SDPBMM which brings almost no degradation in performance. From a series of experiments, it can be concluded that the SDPBMM has the following advantages: 1) It is a simple but effective acoustic modeling method for dialectal speech given only a small amount of data; 2) It can make a significant performance improvement for dialectal speech recognition with almost no degradation for *Putonghua* recognition; 3) It can work well with other existing adaptation methods to further improve the performance on dialectal Chinese.

7. References

[1] S.Goronzy, R. Kompe, S. Rapp, "Generating Non-Native Pronunciation Variants for Lexicon Adaptation," *Speech Communication*, Vol. 42(1), pp.109-123, 2004.

[2] Y. Liu, P. Fung, "Pronunciation Modeling for Spontaneous Mandarin Speech Recognition," *International Journal of Speech Technology*, Vol. 7, pp.155-172, 2004.

[3] L.M. Tomokiyo, "Recognizing Non-Native Speech: Characterizing and Adapting to Non-Native Usage in LVCSR," *PhD Thesis, Carnegie Mellon University*, 2001.

[4] J. Li, T.F. Zheng, W. Byrne, D. Jurafsky, "A Dialectal Chinese Speech Recognition Framework," *Journal of Computer Science and Technology*, Vol. 21(1), pp.106-115, 2006.

[5] C. Huang, T. Chen, E. Chang, "Accent Issue in Large Vocabulary Continuous Speech Recognition," *International Journal of Speech Technology*, Vol.7, pp. 141-153, 2004.

[6] Y.L. Zheng, R. Sproat, L. Gu *et al.*, "Accent Detection and Speech Recognition for Shanghai-Accented Mandarin," *Interspeech*, Lisbon, 2005.

[7] M.Y. Hwang, X.-D. Huang, F.A. Alleva, "Predicting Unseen Triphones with Senones," *IEEE Transaction on Speech and Audio Processing*, Vol.4(6), pp. 412-419, November, 1996.

[8] E.F. Lussier, "A Tutorial on Pronunciation Modeling for Large Vocabulary Speech Recognition," *Lecture Notes in Computer Science*, Springer Berlin/Heidelberg, Vol.2705, pp.38-77, 2003.

[9] R. Gruhn, K. Markov, S. Nakamura, "A Statistical Lexicon for Non-Native Speech Recognition," *ICSLP*, Korea, 2004.

[10] S. Young, G. Evermann, T. Hain, *et al.*, "The HTK Book (for HTK Version 3.2.1)," *Cambridge University*, Cambridge, 2002.

[11] A.J. Li, X. Wang, "A Contrastive Investigation of Standard Mandarin and Accented Mandarin," *EuroSpeech*, Geneva, 2003.

[12] Z.R. Wang, T. Schultz, A. Waibel, "Comparison of Acoustic Model Adaptation Techniques on Non-native Speech," *IEEE ICASSP*, pp.540-543, Hong Kong, 2003.

[13] L.Q. Liu, T.F. Zheng, W.H. Wu, "State-Dependent Phoneme-Based Model Merging for Dialectal Chinese Speech Recognition", *ISCSLP*, Singapore, 2006.