Automatic Initial/Final Generation 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, China
liulq@cst.cs.tsinghua.edu.cn, fzheng@tsinghua.edu.cn, wuwh@tsinghua.edu.cn

ABSTRACT

Phonetic differences always exist between any Chinese dialect and standard Chinese (*Putonghua*). In this paper, a method, named automatic dialect-specific Initial/Final (IF) generation, is proposed to deal with the issue of phonemic difference which can automatically produce the dialect-specific units based on model distance measure. A dialect-specific decision tree regrowing method is also proposed to cope with the tri-IF expansion due to the introduction of dialect-specific IFs (DIFs). In combination with a certain adaptation technique, the proposed methods can achieve a syllable error rate (SER) reduction of 18.5% for Shanghai-accented Chinese compared with the *Putonghua*-based baseline while the use of the DIF set only can lead to an SER reduction of 5.5%.

Index Terms: speech recognition, dialectal Chinese speech recognition, phone set generation, acoustic distance measure

1. INTRODUCTION

Dialect is a language spoken by people living in a certain region. In China, standard Chinese, or *Putonghua*, is the official dialect through which Chinese people from different regions are mutually understandable. However when Chinese people speak *Putonghua*, they are often affected by their native dialects to some extent. In our study, we refer to *Putonghua* influenced by a certain native dialect as dialectal Chinese (DC). For dialectal Chinese speech recognition, most of the state-of-the-art automatic speech recognition (ASR) systems fail to perform well. On average, an absolute performance degradation of 10~30% always occurs dependent on the speaker's accent level, or in other words the level of mismatch between training and test speech. To deal with the mismatch acoustically, a number of mechanisms have been proposed in some practical systems: 1) Bootstrapping, the most straightforward approach, which retrains the dialect-specific model using a standard language as a seed in combination with some dialectal speech data[1]; 2) Acoustic modeling with state sharing, in which phonetic models share the state parameters derived from standard speech as well as dialectal speech [2]; 3) Acoustic model interpolation, which interpolates the standard speech model with the dialect-specific one [3]; 4) Dialect adaptation, which modifies the parameters in the acoustic model using adaptation data [4]. All but the last method mentioned above need a quite large amount of dialectal speech data to train/retrain a robust model for dialectal speech recognition, and hence is too costly in practice. However for dialect adaptation, a relatively small amount of dialectal speech data is sufficient to obtain a robust model for dialectal speech recognition [1].

So far as a dialect is concerned, the phoneme set is definitely not the same as that of *Putonghua*. For example, in *Putonghua*, there are 21 Initials and 38 Finals; however, there are 54 Finals, 34 Initials in Shanghai dialect [5]. How to capture the differences between a dialectal Chinese and *Putonghua* in the phonetic level is worthy of investigation. The idea comes from the assumption that the widely used IF set for *Putonghua* cannot be used to characterize the dialectal Chinese precisely because there are some dialect-specific units which can not be represented by the standard IFs. In [6], 13 dialect-specific IFs were selected from the surface form transcription in terms of their frequencies where the proposed approach relies heavily on the detailed transcription. Hereinafter, an IF set consisting of both the standard IFs and the dialect-specific IFs is referred to as a DIF set. In our study, an approach to automatic generation of DIFs is proposed. Two issues should be dealt with properly: 1) What criterion should be used to generate the dialect-specific IFs; 2) How to deal with the context-dependent HMMs expansion which, without proper handling, will most likely worsen the ASR performance. For the first issue, to recognize dialectal Chinese effectively based on a *Putonghua* model and a small amount of dialectal Chinese speech data, a simple but effective data-driven method for automatic dialect-specific IF generation (ADIFG) is proposed. Based on the minimum model distance, the ADIFG calculates the model distances between any standard IF's HMM and its corresponding dialect-specific HMM in order to generate the DIF set. Considering that the dialect-specific IFs are closely related to their corresponding standard IFs, the dialect-specific ones can be regarded as the variants of their corresponding standard phonemic units affected by a certain dialect. For the second issue, a clustering method namely dialect-specific decision tree regrowing (DDTR) is proposed which is inspired by polyphone decision tree specialization (PDTS) [7]. One motivation here is to build a robust recognizer for a dialectal Chinese based on a handy *Putonghua* model with a small amount of dialectal speech data. Taking Shanghai dialect as the target experimental dialect, we adopted the adaptation method, maximum likelihood linear regresssion (MLLR), as well as DIFs and DDTR to improve the performance for Shanghai dialectal Chinese speech recognition. As a result, a relative syllable error rate (SER) reduction of 18.5% was achieved in comparison with the *Putonghua*-based HMMs.

The remainder of this paper is organized as follows. The details of the proposed ADIFG are described in Section 2 with two issues discussed concerning the DIF generation, namely, the DIF generation criterion and the DDTR. In Section 3, the experiments based on the DIF set and the DDTR together with

the MLLR are designed and performed to verify the effectiveness of the proposed methods. Finally, conclusions are drawn in Section 4.

2. AUTOMATIC DIALECT-SPECIFIC INITIAL/FINAL GENERATION

2.1 The standard and the dialect-specific IFs

One of the typical characteristics of the Chinese language is the IF structure. Almost any Chinese syllable (Pinyin) consists of an Initial and a Final. For a large vocabulary continuous speech recognition system, Initials and Finals are commonly used as speech recognition units. In *Putonghua*, there are 21 Initials and 38 Finals. For a certain dialect, the set will be different. For example, in Shanghai dialect, there are 34 Initials and 54 Finals. When a native Shanghai speaker speaks *Putonghua*, the pronunciation will be affected by his/her native dialect to some extent. It is safely assumed that the IF set for a dialectal Chinese is a superset of that of *Putonghua*. That is to say, some dialect-specific units should be additionally considered so as to build a robust dialectal-Chinese recognizer. For example, the Final '*ie*' is often pronounced in Shanghai-accented *Putonghua* as '*ie*^' or '*ie*<'. It is true that not all the IFs in a dialect will be definitely seen in the corresponding dialectal Chinese. Additionally, the dialect-specific IFs are not independent of their standard IFs, for example, '*ie*', '*ie*^' and '*ie*>' are closely relevant while some slight difference does exist among them acoustically. Therefore, for dialectal Chinese speech recognition, dialect-specific units should be selected under a certain criterion and then added to the standard IF set to form a rich DIF set.

It is expected that the discriminative ability of models is improved due to adoption of more precise DIFs; but more data is needed to build robust HMMs related to the introduced dialect-specific units. Two problems arising with the expansion of DIFs should be well solved. 1) How to acquire the dialect-specific units? 2) How to deal with the expansion of tri-IFs?

2.2 Criterion for DIF generation

Suppose that A is an HMM for an IF built on *Putonghua* and A' is the HMM for the same IF built on a certain dialectal Chinese. It could be assumed that Model A and Model A' are close to each other acoustically. The closer they are, the more similar the two models are. Likewise, two less similar models will have a bigger distance acoustically. Thus the distance between the *Putonghua* model and its corresponding dialect-specific model can be measured quantitatively. The distance between HMM of a standard IF and one of its corresponding dialect-specific IFs (referred to as *a pair*) is measured with Equation 1 [8]. For simplicity, in the distance measure only the mono-IFs are considered. The dialect-specific mono-IF HMMs can be robustly built based on a relatively small amount of dialectal Chinese. Then the distances across all the IF pairs are averaged so as to define a threshold, *Threshold*. For those IF pairs whose distances are less than *Threshold*, in which case the two models can be regarded as a similar pair, no alterative dialect-specific IF is necessary to be generated. Likewise, for medium- or big-distance pairs, 1 or 2 dialect-specific IFs should be generated. The basic idea for the classification is illustrated in Equation 2 where d stands for model distance.

$$d(i,j)=\frac{1}{M_s}\sum_{s=1}^{M_s}\frac{1}{V_s}\sum_{k=1}^{V}\left[\frac{\left(\mu_{isk}-\mu_{jsk}\right)^2}{\sigma_{isk}\sigma_{jsk}}\right]^{1/2} \quad (1)$$

$$\#\ of\ DIFs=\begin{cases}0, & d\le Threshold\\ 1, & Threshold<d\le 1.5\times Threshold\\ 2, & d>1.5\times Threshold\end{cases} \quad (2)$$

A way to obtain the DIF transcription could be based on the building of HMMs for the newly-generated dialect-specific IFs. The initial dialect-specific mono-IF HMMs are to be generated this way: 1) For a standard IF with one alternative dialect-specific IF, the initial model for the alternative is copied from the corresponding DC HMM; 2) For a standard IF with two alternatives, one is handled the same way as in 1) while the other one is to be derived from the interpolation of the standard mono-IF and its corresponding dialect-specific mono-IF, as illustrated in Equation 3, where λ is a linear interpolation coefficient between the *Putonghua* and the DC acoustic models. The method is based on the assumption that due to a big distance between two models, a new model by means of the interpolation of two models is generated to narrow the gap so that a much bigger space can be covered acoustically.

$$p'\left(x\left|a_{a\_d}\right.\right)=\lambda\, p\left(x\left|a_s\right.\right)+\left(1-\lambda\right)p\left(x\left|a_d\right.\right) \quad (3)$$

In Equation 3, $p'(x|a_{b\_d})$ stands for the interpolated model, $p(x|a_s)$ the standard mono-IF, while $p(x|a_d)$ the dialect-specific mono-IF. The interpolation coefficient λ is normally selected empirically, for example in our experiment λ=0.72 was set experimentally. The Viterbi-based forced-alignment was performed to obtain the DIF transcription for dialectal Chinese speech data. The *Pinyin* layer transcription is sufficient to carry out the forced-alignment and hence the detailed IF layer transcription is unnecessary. The whole process is depicted in Figure 1.

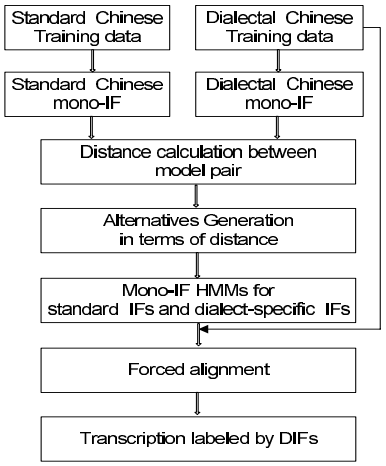

Figure 1: flowchart for automatic DIF generation

2.3 Dialect-specific decision tree regrowing

By adopting the DIF, many initially unseen tri-IFs without any corresponding HMMs are introduced, which is called *tri-IF expansion* in this paper. To solve this problem, a dialect-specific decision tree regowing method is proposed which is inspired by PDTR. In the approach, the clustered *Putonghua* tri-IF decision tree is shared with the dialectal Chinese by restarting the decision tree growing process. To integrate the dialect-specific tri-IFs into the decision tree, some questions concerning these units should be added. During the regrowing process, a node needs splitting if and only if there is sufficient adaptation data available for it in dialectal Chinese. Figure 2 illustrates one part of the decision tree for the second state of standard IF before DDTR while Figure 3 for the dialect-specific decision tree after DDTR. In these figures, *an* is a standard Final, and *an*1 is an *an*-derived dialect-specific Final. Mostly *an*1-centered tri-IFs are of the same hierarchy as *an*-centered ones except that sufficient data is available for some *an*1-centered tri-IFs, in which case a node split will occur. In these figures, newly generated child nodes are labeled with light yellow shadowed circles. Only dialect-specific tri-IFs with sufficient adaptation data are to be produced, those without sufficient adaptation data will share the model parameters with their corresponding standard tri-IFs.

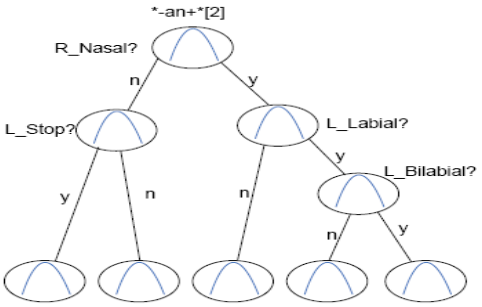

Figure 2: standard IF decision tree

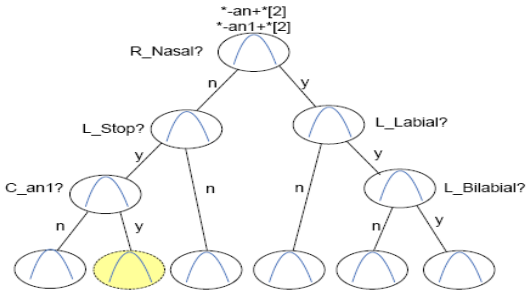

Figure 3: DIF decision tree regrowing

3. EXPERIMENT AND RESULTS

3.1Baseline

The Mandarin Broadcast News (MBN) database, a read style standard Chinese speech corpus, was used to train the baseline *Putonghua* acoustic model. It contained about 30 hours of high quality wideband speech with detailed Chinese IF transcriptions. The acoustic models of *Putonghua*-based baseline were the state-clustered 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 ΔΔ were used as the features with cepstral mean normalization. The model showed good performance statistically upon which many research was carried out [10]. Additionally, 6 zero-Initials were added to the standard IF set to help improve the performance and make the modeling process consistent [6].

Another database, named Wu dialectal Chinese corpus (WDC), contained 100 native Shanghai speakers, 50 males and 50 females. The speech was recorded under a similar condition to that of MBN. The use of this database was to minimize the channel affect. Further details on the database can be found in [9]. The Chinese *Pinyin* level transcription was used to obtain the IF layer transcription via a lookup standard Pinyin-to-IF lexicon. 20 speakers' data from WDC with 970 sentences was used as development set while another 20 speakers' data with 995 sentences as a test set. The test set was composed of speech from medium and strong accented speakers in balance. For comparison purpose, another acoustic model, named *rebuilt DC model*, were rebuilt based on the WDC training data in the same way as that for the *Putonghua* model. The training data for rebuilding the DC model contained 80 speakers with totally 3,860 sentences. Because acoustic modeling was the research focus no language models were used. HTK [8] was used in the experiments. The details are presented in Table 1.

Table 1: details for experimental setup

Items	Baseline (*Putonghua*)	Rebuilt DC model
Number of states	3,230	3,172
Number of mixtures	45,220	44,408
Number of IFs	65	65
Number of tri-IFs	7,411	7,174
Features	39 MFCC+ Δ, ΔΔ, CMN	
Dictionary	406 Chinese Pinyin loop network	
Training data	MBN/30 hours	WDC/4 hours
Test data	20 Shanghai-accented speakers, approximately 1 hour, 995 utterances	
SER	49.3%	38.1%

The *Putonghua* model achieved an SER of only 49.3% primarily due to the Shanghai accent. The dialectal Chinese model built using only dialectal speech achieved an SER of 38.1%, which to some extent can be regarded as the upper bound for dialectal Chinese recognition by means of approaches adopted presently.

3.2 MLLR-based adaptation with DIF

As for context-independent acoustic modeling, it does not need so much training data as context-dependent modeling does. Thus, the development set taken from WDC was used to train the dialect-specific mono-IF HMMs. The model distance between a certain *Putonghua* and its corresponding DC model was measured by Equation 1. According to the ADIFG, there

were 24 alternatives produced finally. As a compromise between data availability and discriminative ability, there were 16 IFs each having 1 alternative, and 4 Finals each having 2 alternatives. Consequently, there were 89 IFs, including 65 standard IFs and 24 Shanghai dialect-specific alternatives.

Along with the adaptation technique, speaker-independent models can be transformed to compensate for differences in the acoustic environment or the characteristic of a group of speakers. Thus, speakers of a same dialect are good candidates for adaptation because of the consistency of many deviations from the *Putonghua* pronunciation. Most widely used adaptation techniques include the MLLR and the maximum *a posteriori* (MAP) adaptation methods. 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 DIF transcription was obtained in the manner mentioned in Section 2.2. All DIFs were classified into 89 classes, with each corresponding to an IF; both the diagonal and the bias offsets were used in the MLLR transformation matrix. The development set from WDC was used as adaptation data. The result is given in Figure 4.

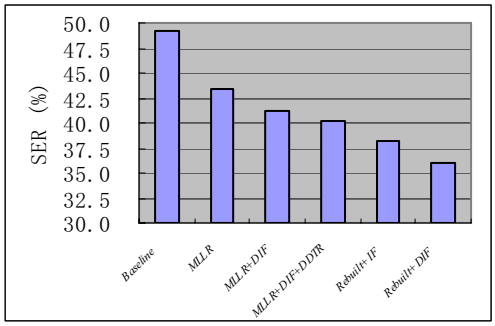

Figure 4: results for dialect-specific IF generation

It is shown by column *MLLR* that with the application of MLLR together with the standard IFs, an SER of 43.4% was achieved with an SER reduction of 12.0% compared with the baseline. When DIFs were adopted together with the MLLR adaptation, another drop in SER of 2.3% absolutely was brought, as depicted by column *MLLR+DIF*. In total, an SER reduction of 16.6% was achieved when adopting the MLLR and the DIF set simultaneously. A conclusion could be drawn experimentally that the DIF could really improve the accuracy for DC speech recognition. The results of another supportive experiment are also given in Figure 4. Namely, the rebuilt acoustic model with DIF set, column *Rebuilt+DIF,* used only DC speech data, which was also the same training set for column *Rebuilt+IF*. Accordingly, an SER of 36.0% or an SER reduction of 5.5% was achieved.

3.3 Effectiveness of DDTR

Prior to the adaptation with DIF from standard IF HMMs, DDTR was applied with the goal of building more robust HMMs for dialectal Chinese. The effectiveness is also presented in Figure 4. From column *MLLR+DIF+DDTR*, we can see that a further absolute SER reduction of 0.9% was achieved, which verifies that the introduction of dialect-specific IFs is of two sides: on the one hand, it can improve the discriminative ability of HMMs; on the other hand, some redundant parameters do exist such that the sharing mechanism is a necessity. It is shown experimentally that DDTR is a good way to make the balance. In summary, an overall relative SER reduction of 18.5% was obtained compared with the baseline.

4. CONCLUSION

In this paper, we report the approach to the generation of DIFs automatically based on the model distance measure between the *Putonghua* mono-IF HMMs and the dialect-specific mono-IF HMMs trained with a small amount of dialectal Chinese speech. The combination of the MLLR with the DIF set could achieve an SER reduction of 16.6% for Shanghai-dialectal Chinese speech recognition compared with the baseline. An SER reduction of 5.5% could be achieved via the DIF set when rebuilding the DC model with dialectal Chinese speech data. If using the DDTR, proposed to build the context-dependent HMMs for the introduced dialect-specific IFs, as well as the MLLR and the DIF set, another absolute SER reduction of 0.9% could be achieved. It can be seen that the best solution is *MLLR+DIF+DDTR* which can lead to an SER reduction of 18.5% for Shanghai-dialectal Chinese speech recognition. Experiments in this paper clearly show that the dialect-specific IF generation method together with a certain adaptation technique can improve the performance for dialectal speech recognition significantly.

5. REFERENCES

[1] Diakoloukas,V., Digalakis, V., Neumeyer, L. and Kaja, J.," Development of Dialect-Specific Speech Recognizers Using Adaptation Methods", *IEEE ICASSP*, 1997.

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

[3] Livescu, K., "Analysis and Modeling of Non-Native Speech for Automatic Speech Recognition", Master thesis, *MIT*, 1999.

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

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

[6] Li, J., Zheng, T.-F., Byrne, W. *et al.* "A Dialectal Chinese Speech Recognition Framework", *Journal of Computer Science and Technology*, 21(1): 106-115, Jan. 2006.

[7] Wang, Z.-R., and Schultz, T., "Non-Native Spontaneous Speech Recognition through Polyphone Decision Tree Specialization", *EuroSpeech*, 2003, Geneva.

[8] 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/.

[9] Li, J., Zheng, F., Xiong, Z.-Y. *et al.* "Construction of Large-Scale Shanghai Putonghua Speech Corpus for Chinese Speech Recognition", *Oriental-COCOSDA*, 62-69, October, 2003, Singapore.

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