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
liul@cst.cs.tsinghua.edu.cn, fzheng@tsinghua.edu.cn, wuhw@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 cannot 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 regression (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, if two dialect-specific IFs are close enough, they are assumed to be in the same dialect. In Figure 1 the two models are. Likewise, two less similar models will be farther apart acoustically. The closer they are, the more similar the distance between two models, a new model by means of the interpolation of the standard IFs and its corresponding dialect-specific IFs, as illustrated in Figure 3, where \( \lambda \) 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 the standard IFs and its corresponding dialect-specific IFs, as illustrated in Figure 3, where \( \lambda \) is a linear interpolation coefficient between the Putonghua and the DC acoustic models.

\[
d(i,j) = \frac{1}{M_x} \sum_{t=1}^{M_x} \sum_{k=1}^{K} \frac{(\mu_{ik} - \mu_{jk})^2}{\sigma_{ik}^2 \sigma_{jk}^2}
\]

\( (i,j) \) the standard mono-IF, while \( x|a_d \) the dialect-specific mono-IF, as illustrated in Equation 3, where \( \lambda \) is normally selected empirically, for example in our experiment \( \lambda=0.72 \) 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.

\[
p \left( x | a_{u,a} \right) = \lambda p \left( x | a_u \right) + (1 - \lambda) p \left( x | a_a \right)
\]

\( \lambda \) is the dialect-specific IF. The interpolation coefficient \( \lambda \) is normally selected empirically, for example in our experiment \( \lambda=0.72 \) 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 regrowing 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 an1 is an an-derived dialect-specific Final. Mostly an1-centered tri-IFs are of the same hierarchy as an-centered ones except that sufficient data is available for some an1-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](image1)

![Figure 3: DIF decision tree regrowing](image2)

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

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](image)

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 dialect 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