DIALECT ADAPTATION FOR MANDARIN CHINESE SPEECH RECOGNITION

Frédéric Beaugendre, Tom Claes, Hugo Van hamme
Lernout and Hauspie, Koning Albert I laan 64, B-1780, Wemmel, Belgium
Email: frederic.beaugendre@lhs.be

1. INTRODUCTION

Recent advances in speech recognition research have resulted in high performance speaker independent (SI) recognition systems. Such systems perform well under the constraint of being trained on very large amounts of speech data. However, even a considerable amount of speech data will not cover all the possible speech and channel variability, and with good SI systems, some speakers are still modeled poorly, resulting in poor accuracy.

Since it is not practical to collect a large amount of speech data for a specific speaker and retrain the models from scratch, the problem can be solved by adapting the SI models towards speaker dependent (SD) models, using much smaller amounts of speech than those used in conventional training techniques.

The most recent adaptation techniques used for large vocabulary speech recognition using Gaussian mixture models include transformation based methods [2, 3], Bayesian adaptation [4, 5] or a combination of both techniques [1].

Recent studies have also shown that those adaptation techniques can be generalized to adapt a specific dialect dependent (DD) SI model to an other dialect dependent (D’D) SI model [6].

Such techniques are of great interest for Mandarin Chinese, where wide variety of local or regional dialects exist. In this paper, we therefore investigate the development of a dialect-specific recognition system in Mandarin Chinese using standard adaptation techniques. Lexical adaptation is not considered here.

The paper is organized as follow: section 2 describes the adaptation methods investigated in this study. Section 3 presents the experimental set-up and summarizes the results obtained. Section 4 concludes.

2. ADAPTATION METHODS

In the present work, the SI models of the source dialect use an HMM with tied distributions modeled by a mixture of Gaussians drawn from a large pool, and shared across all distributions. The system's front-end uses a configuration of 12 MFCC and energy with their first and second order derivatives.

For tone recognition, a separate discrete density model is trained using normalized pitch and energy features augmented with their derivatives. During recognition, the continuous and discrete density log-probabilities are linearly combined.

The most widely used techniques proposed to improve the speech recognition performance in mismatched conditions include transformation-based adaptation (MLLR), Bayesian adaptation or combined approaches. Those techniques will be considered in our study. However, for a detailed description of the implementation of those methods, the reader is referred to [3] and [1] respectively. In addition, a weight interpolation mechanism described in 2.3 is added in order to further improve the performance. Finally, tone adaptation is also considered.

2.1. Transformation based adaptation

A technique applicable to both speaker and environment adaptation is maximum likelihood linear regression (MLLR). The method was originally designed to estimate a set of linear transformations for the mean parameters of a mixture Gaussian HMM system to maximize the likelihood of the adaptation data. These transformations can capture general relationships between the original model set and the current speaker or new acoustic environment.

In an initial study, the authors confirm that the other HMM parameters do not need to be adapted since the main differences between speakers are assumed to be characterized by the means [3]. In later development of this work, it has been shown that the model variances could be updated within the same framework, which allows improvements in case of environment mismatches [5]. As the databases used in this study were recorded in the same conditions, the implementation of the MLLR transformation is limited to a single full transformation matrix to transform all the Gaussian means.

Figure 1a presents the diagram of this experiment.

2.2. Combined adaptation

A second family of adaptation algorithms follows a Bayesian approach, where the speaker-independent information is encapsulated in the prior distributions. The Bayesian approach has the nice property that the adapted model performance will converge to the target model performance as the amount of adaptation speech increases. However, the adaptation rate is usually slow. Bayesian adaptation works well for large amounts of data, while our implementation of the transformation based approach work for small amounts of data but don’t converge to
the ML model trained on D’ only, as the amount of data increases. The idea of combining the 2 methods is that the prior model in Bayesian adaptation can be improved by replacing it with a model estimated with a transformation method [1].

Figure 1b presents the diagram of this adaptation technique in the framework of the present study.

**Figure 1:** Presentation diagram of HMM adaptation using MLLR (a), Combined (MLLR+MAP) (b) techniques (adapted from [1]). DD = dialect dependent, DA = dialect adapted, SI = Speaker independent.

### 2.3. Weight interpolation

An additional technique was also implemented in order to find an estimate of the dialect dependent weights, by interpolating between the source dialect dependent and intermediate dialect adapted weights, using an interpolation factor dependent on the amount of adaptation data used:

\[ \tilde{W}_{dd'} = W_{dd'} + \log(1 + \alpha n) W_{dd} \]  

(1)

with

- \( W_{dd'} \cdot W_{dd} \) normalized weights
- \( n \) : total number of frames for the corresponding \( W_{dd} \)
- \( \alpha \) : control parameter

The parameter \( \alpha \) needs to be tuned in order to get a good balance between dialect dependent and full SI model weights.

### 2.4. Tone adaptation

A weighting function between the counts for the discrete codebooks of the source and intermediate adapted target models could be investigated in case of tone mismatch between dialects.

However, reference experiments carried out in order to investigate the need for tone adaptation across the pair of dialects studied have shown that this adaptation is not useful (see 3.6).

### 3. EXPERIMENTS DESIGN AND RESULTS

#### 3.1. Database

The experiments reported in this report are based on two validated databases containing microphone speech recorded in different regions of China. Around 100 speakers were selected for each dialect (region of Beijing and Taiwan) for a total of around 80 hours of speech. The corpus consists of subjects reading various prompts organized in sections. Each section contains a set of syllables, commands, words, or newspaper sentences.

In order to conduct our research, the two databases were divided into four data sets:

- training data for DD SI source models (Beijing dialect): contains approximately 80 hours of speech data from speakers native from the region of Beijing;
- training data for D’D SI target models (Taiwan dialect): baseline continuous density source models were trained in order to get a performance reference for the adapted models. The amount of speech data is comparable to what is used for the source models;
- adaptation data (Taiwan dialect): several subsets of the target dialect training database (ranging from 10 to 2500 utterances) were used for adaptation;
- test data (Taiwan dialect): all recognition experiments presented in this report use this data set; it contains data of 10 speakers that consist of 3799 utterances or 3h19min of speech, and is fully independent from the training and adaptation sets.

#### 3.2. Baseline models

Prior to any adaptation experiment, baseline models were built for the source and target dialects using the two first data sets described in the previous section. A reference test experiment was conducted using those two models.

The reference word error rate obtained (before adaptation) are 28.3% using the source dialect models, and 7.1% and 5.9% using the target dialect models, without and with tone integration respectively. The important difference observed between the two reference models attest for the potential for improvement using adaptation methods. In addition, the degradation in performance obtained for the source models was uniform across speakers, which indicates consistent differences between the two dialects.

#### 3.3. MLLR adaptation

In the first set of experiments, the MLLR scheme was used to adapt the source models using various size of adaptation data sets, ranging from 10 to 250 utterances.

An evaluation of the MLLR-adapted models is achieved using the test set described in section 3.1. Figure 2 and table 1...
summarizes the results. We can observe that the average WER improve significantly from 28.3% to 23.9%, which gives 15.5% relative WER reduction. In addition, the comparison of the improvement obtained per speaker show significant differences across speakers (from 7.6% to 29.3% relative).

The relatively small improvement compared to MLLR performance obtained in speaker adaptation tasks may be explained by the fact that the global transformation applied may be more biased towards one particular speaker and lead to worse performance on another speaker [2].

In [6], MLLR performance on Swedish dialect adaptation using a similar transformation procedure gives a relative reduction of the WER of 38% (from 25% to 13% using 500 utterances). However, the difference with the results obtained for Chinese can be explained by differences relative to the language, the test set used, the number of adaptation sentences, or a combination of these parameters.

Finally, as both data sets contain clean speech with a relatively low background noise, no significant improvement is obtained by including the silence states for adaptation.

Figure 2: Performance results (WER in %) using a full transformation matrix MLLR adaptation scheme from the source models.

3.4. Combined adaptation

This second adaptation experiment was conducted with a combined adaptation scheme repeated for different adaptation data sets of 75, 150, 250, 500, 1000, 1500 and 2500 utterances.

Performance results of the Combined-adapted models are given in figure 3 and table 1. Experiments were repeated with and without silence adaptation, with strictly identical results. The WER obtained is 15% with 500 utterances used for adaptation, which gives a relative reduction of 46.6%. As expected, the WER continues to decrease as the number of adaptation utterances increases (we reach 12.8% WER, i.e. a reduction of 54.8% relative for 2500 utterances). However, the asymptotical value towards which the performance converges is significantly higher than the one obtained with the target models. This can easily be explained by the fact that only the Gaussian means are adapted in the present framework of the adaptation scheme.

3.5. Weight interpolation

In a first stage, the standard adaptation techniques mentioned above were used to adapt the Gaussian means. In a second stage, the target dialect model weights were also adapted using the interpolation technique mentioned in 2.3.

After an initial tuning experiment, an optimal value of 1 was found for the parameter $\alpha$.

Mixture weight adaptation was applied to models obtained from the Combined adaptation. Adaptation results for the different data sets are given in table 1 and figure 3. Significant improvement is observed: for 500 utterances used, the WER obtained is 13.9% (8% relative improvement from Combined model performance and 54.8% from source models); for 2500 utterances, WER is then 11.0% (14.1% relative improvement from Combined model performance and 61.1% from source models).

3.6. Tone adaptation

The experiments presented from the beginning of this section were carried out using the standard continuous density engine architecture of our system. In this context, the tones were only indirectly taken into account given the fact that we are using tone dependent speech units (for FINALS) as reference.

In this last experiment, we used a multi-stream recognizer combining the probabilities provided by the continuous emissions using the standard feature stream and discrete emissions using a discrete tone codebook. The late introduction of this mechanism in this work was done on purpose in order to clearly separate the contribution of standard mixture Gaussian adaptation, and multi-emission mechanism and tone adaptation.

As it has been done for the standard continuous density engine, reference tone models were trained using the target dialect training set in order to get a baseline performance for tone adapted models. Moreover, an initial tuning experiment of the multi-emission recognizer was necessary to optimize the weights assigned to the contribution of the continuous and discrete emissions. The WER obtained for the optimal weight parameter is 5.9%, which gives a reduction of 16.9% relative from the performance using the standard continuous density engine architecture (7.1% WER).
A test was achieved using the tone codebook trained on the target dialect, and the continuous density models obtained after combined adaptation and weight interpolation. The multi-stream recognizer gives a WER of 10.6%, which is a relative improvement of 23.7% from the performance using the same models without tone mechanism, and using 500 utterances for adaptation. The relative improvement is 26.5% using 75 utterances, and 21.8% using 2500 utterances.

No significant improvement or degradation is observed using the tone codebook trained on the source dialect. This allows us to conclude that no dialect adaptation is needed for the tone codebook.

<table>
<thead>
<tr>
<th>Nb of adapt. sentences</th>
<th>10</th>
<th>250</th>
<th>500</th>
<th>2500</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLLR adapt.</td>
<td>24.3</td>
<td>24.3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Combined adapt.</td>
<td>-</td>
<td>16.5</td>
<td>15.1</td>
<td>12.8</td>
</tr>
<tr>
<td>idem + weight interp.</td>
<td>-</td>
<td>15.3</td>
<td>13.9</td>
<td>11.0</td>
</tr>
<tr>
<td>idem + tone integration (tone model trained on source dialect)</td>
<td>-</td>
<td>11.6</td>
<td>10.9</td>
<td>8.7</td>
</tr>
<tr>
<td>idem + tone integration (tone model trained on target dialect)</td>
<td>-</td>
<td>11.6</td>
<td>10.6</td>
<td>8.7</td>
</tr>
</tbody>
</table>

Table 1: Word Error Rate (in %) using different adaptation techniques and adaptation sets.

4. CONCLUSION

This paper summarizes a series of experiments carried out in the framework of adaptation of continuous density models to Chinese regional dialects.

In a preliminary comparison of the two dialects studied, we observed an important degradation of the performance in case of dialect mismatch.

Even if a significant improvement is obtained using the MLLR adaptation scheme, the method shows clear limitation compared to the performance obtained on speaker adaptation. However, combined with a Bayesian adaptation and a weight interpolation technique, dramatic improvement of the performance of our system is obtained, which lead us to the conclusion that dialect adaptation is a very interesting alternative to standard ML training in terms of trainability.

Finally, the adaptation of tone recognition was also investigated. In an initial experiment, a reference test was conducted using adapted and non-adapted tone models. No improvement is obtained using tone models trained on the target dialect, which turns tone adaptation to be useless.

5. ACKNOWLEDGMENTS

This work was supported by the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT).

6. REFERENCES