Nonnative Speech Recognition Based on State-Candidate Bilingual Model Modification
Qingqing Zhang, Ta Li, Jielin Pan and Yonghong Yan
ThinkIT Speech Laboratory
Institute of Acoustics Chinese Academy of Sciences
Beijing 100080, China
zhangqqingqing, lita, panjielin, yanyonghong@hcc1.ioa.ac.cn

Abstract
The speech recognition accuracy has been observed to decrease for nonnative speakers, especially those who are just beginning to learn foreign language or who have heavy accents. This paper presents a novel bilingual model modification approach to improve nonnative speech recognition, considering these great variations of accented pronunciations. Each state of the baseline nonnative acoustic models is modified with several candidate states from the auxiliary acoustic models, which are trained by speakers’ mother language. State mapping criterion and n-best candidates are investigated based on a grammar-constrained speech recognition system. Using the state-candidate bilingual model modification approach, compared to the nonnative acoustic models which have already been well trained by adaptation technique MAP, a Relative reduction of 7.87% in Phrase Error Rate (RPhER) was further achieved.

Index Terms: Nonnative speech recognition, bilingual model modification, pronunciation variation

1. Introduction
With globalization, nonnative speech recognition has become a popular issue in ASR. As opposed to native one, current speech recognition is known to perform considerably worse when recognizing nonnative speech, which results in word error rates of two to three times native error rates [1]. The problem of recognition of nonnative speech results from mismatches between acoustic models automatically determined from training data and characteristics of test data. Usually, acoustic models are trained on native speech, which represents characteristics of native pronunciation. Nonnative speakers’ pronunciations of the test speech, however, different from those native speakers’ observed during training, dramatically decreases the recognition performance [2]. [3] [4] show that when acoustics of the accent are taken into account, large gains in performance can be achieved. This indicates that if acoustic models are trained with the nonnative speech covering the pronunciation variation, the performance of nonnative speech recognition can be improved.

In fact, the variation among nonnative speakers, even with the same accent, is potentially great [5]. The characteristics of the nonnative speakers’ pronunciation may differ in fluencies, levels of familiarity with the target language, individual tendencies in mapping unfamiliar sounds with their own native language sounds, etc. In order to build robust nonnative acoustic models, much more training data with different kinds of varieties are required. However, to obtain such training data is very difficult, especially for those with heavy accents, since most of the nonnative training data are recorded by speakers who are familiar with the second language relatively. In recent years, speaker adaptation techniques such as MAP and MLLR have been widely used to handle nonnative speech [1] [6], by which native acoustic models can be adapted to nonnative ones with small amounts of adaptive speech. In these methods the similarity between adaptive speech and test speech is the key point which determines the performance. Although these techniques gain in nonnative speech recognition accuracy, the typical nonnative accuracy following adaptation still falls substantially below that of native speech. [7] studies the limitations of adaptation. Results suggest that the target language phones that do not exist in the speakers’ own language as well as the great acoustic variability reduces the performance heavily. How to make acoustic models robust and tolerant with these pronunciation variations is the motivation of our research presented in this paper.

Flege et al. [8] argues that nonnative speakers may produce speech sounds which are either part of their own native language or which are established via merging characteristics of a native sound with a nonnative speech sound. Thus it can be speculated that the training data from the speakers’ mother language will be useful for nonnative speech recognition, especially when the heavy accent occurs (where speakers tend to use the phonemes of their own native language to substitute ones of the target language), since the training data represents pronunciation characteristics of the speakers’ own native language.

Motivated by this, we examine the problem of Mandarin-accented English speech recognition when only a small amount of nonnative training data is available. In order to make baseline acoustic models to better adjust to characteristics of Mandarin-accent, the baseline nonnative English acoustic models are modified with Mandarin acoustic models at the state level. The similarity between states from the nonnative English acoustic models and Mandarin acoustic models is investigated, and different numbers of state candidates are compared based on the designed testing database.

The paper is structured as follows: Training and testing databases are presented in section 2. In section 3, we describe the baseline acoustic models of our experiments and in section 4 we document how state-candidate bilingual model modification can help to improve the baseline nonnative recognizer performance. Section 5 gives a brief conclusion of this paper.

2. Database description
Our study is restricted to nonnative English spoken by native speakers of Mandarin. All the speech data are recorded through
telephone lines and digitized at 8 KHz sampling rate with 16-bit resolutions. The speech feature vector consists of 36 components (12 PLP parameters, and their first and second order time derivatives), which is analyzed at a 10msec frame rate with a 25msec window size. Cepstral Mean Subtraction (CMS) is employed.

Our training data are divided into three categories: native Mandarin database, native English database and Mandarin-accented English database. The native Mandarin training database consists of native Chinese speech database of National 863 Hi-Tech Project (DB863). It is a standard database published by governmental research program 863 for read speech in Mandarin. The English training database is Wall Street Journal (WSJ). The Mandarin-accented English database was collected in our lab (labeled as ITLab (ThinkIT Libratory)). It includes speech data from hundreds of Mandarin-accented speakers whose English fluencies are above average level. These speakers were asked to read English news sentence by sentence. If there were some words they did not know, they should pronounce them as best as they can. Table 1 summarizes the main information about these three databases.

<table>
<thead>
<tr>
<th>Training Corpus</th>
<th>Type</th>
<th>Source</th>
<th>Time (hour)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TrainM</td>
<td>Native Mandarin</td>
<td>DB863</td>
<td>500</td>
</tr>
<tr>
<td>TrainE</td>
<td>Native English</td>
<td>WSJ</td>
<td>232</td>
</tr>
<tr>
<td>TrainA</td>
<td>Mandarin-accented English</td>
<td>ITLab</td>
<td>20</td>
</tr>
</tbody>
</table>

In order to study different variations of nonnative pronunciations, a representative testing database is designed in the experiment1. This testing database contains two types of phrases that compose of words with different degrees of familiarity to speakers. Some of these phrases have the most common and simply words in English such as "baby", "hello", so the speakers can pronounce them more correctly; the other include some words that are rarely used in general conversations, such as "Fitzgerald", "orchestra", which are hard for Mandarin speakers to produce, and in this situation speakers tend to pronounce them with heavier accents. This database consists of 1568 utterances in total, which were all recorded by Mandarin residents. The English phone set used in experiments is supplied by the ARPABET and the dictionary is based on CMU pronunciation dictionary [9]. This dictionary consists of approximately 53,000 words with associated phonetic transcriptions. All experiments described in this paper use the grammar-constrained speech recognition system.

### 3. Baseline models

In this section, native English acoustic models and different nonnative English acoustic models are investigated and compared. Acoustic models that have the best performance will be selected to be the baseline and modified by the state-candidate bilingual model modification approach in order to achieve further higher nonnative speech recognition accuracy. At the same time, native Mandarin acoustic models are established in this section, which will be used as the auxiliary acoustic models in the model modification approach.

#### 3.1. Native acoustic models

For the sake of comparison, native English acoustic models are established first. Native English acoustic models are trained on TrainE database only, which is labeled as Model_Nat (short for Native acoustic Models). These acoustic models are state clustered crossword triphone HMMs with 32-component Gaussian mixture output densities per state, with about 6000 states in total. Table 2 gives the Phrase Error Rate (PhrER) of the native English acoustic models.

<table>
<thead>
<tr>
<th>Acoustic Models</th>
<th>PhrER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model_Nat</td>
<td>46.9</td>
</tr>
<tr>
<td>Model_ReT</td>
<td>39.3</td>
</tr>
<tr>
<td>Model_MAP</td>
<td>34.3</td>
</tr>
</tbody>
</table>

Additionally, we use the Mandarin training data TrainM to build native Mandarin acoustic models Model_Mand which will be used in the model modification approach as auxiliary acoustic models. The phone set of this model is supplied by 49 toneless Mandarin phones based on the phonetic inventory of the International Phonetic Association (IPA) [10].

#### 3.2. Nonnative acoustic models

In order to improve nonnative speech recognition accuracy, adding nonnative speech data into training process may be the simplest way to reduce the mismatch between training/testing data. We explored some acoustic modeling methods to add the nonnative speech data TrainA into training for fitting characteristics of Mandarin accent. Table 2 shows performances of these nonnative acoustic models on the testing data and gives comparison with native ones.

Model_ReT (short for ReTrained acoustic Models) is the acoustic models retrained by pooling non-native data TrainA and native data TrainE together. Model_MAP (short for MAP acoustic Models) refers to acoustic models to apply adaptation technique MAP on native acoustic models Model_Nat with TrainA. As expected, both of these nonnative models perform better on the nonnative testing data than do the native models. Compared to Model_ReT, Model_MAP reaches the lower PhrER. Thus Model_MAP is selected to be the baseline nonnative acoustic models, and all the improvements with state-candidate bilingual model modification approach below will be based on this performance for nonnative testing data (34.3%).

### 4. State-candidate model modification

Adaptation technique is shown to be efficient for nonnative speech recognition. This technique depends on the consistency between the types of adaptive data and test data. In practical use, however, the nonnative speech recognizer may encounter speakers who are just beginning to learn the foreign language or who have heavy accents. In these cases, nonnative speakers tend to substitute sounds of their mother language for those foreign sounds they can not produce [11]. Usually, these pronunciation substitutions are difficult to predict and hard to capture.
in the training data. Lack of adaptive data, the gain achieved by adaptation technique will be limited.

In our application, nonnative speakers may produce English sounds as part of Mandarin sounds. It can be speculated that modifying nonnative English acoustic models from Mandarin ones may be useful to capture the pronunciation substitutions. Since Mandarin have different phone sets from English, in order to implement the modification, a state-candidate mapping to define which English model state combines with which Mandarin model states is required. In the following section, State-candidate mapping algorithm is presented first, and different n-best state candidates are investigated and compared based on the nonnative test set.

4.1. State-candidate mapping algorithm

State-candidate mapping algorithm used in the paper is an extension of phone clustering algorithm based on confusion matrix [2]. In the state-candidate mapping, each baseline acoustic model state is modified with its corresponding auxiliary model states by combining the Gaussian mixtures together. In our application, nonnative English acoustic models are chosen to be the baseline acoustic models and Mandarin acoustic models are used as auxiliary acoustic models. The detailed algorithm can be described as below: (For convenience, Mandarin and English are referred as source and target language respectively.)

1. Target reference: Force align target language states in small amounts of target language speech data using target language acoustic model in order to get the time-label information. These are considered as the target states reference.

2. Source hypothesis: The source language state recognizer is applied using source language acoustic model on these speech data to decode the source phonetic representation of each utterance. This yields parallel phonetic segmentation of the target language speech data in the source language state inventories. The source phonetic representation is considered as the source states hypothesis.

3. Co-occurrence criterion: Define a criterion for co-occurrence between two phonetic labeling of the reference and hypothesis. In our system, when the number of overlapping frames between the reference and hypothesis states is more than sixty percent of the reference state duration, we can arrange the state of the source language into a target language matrix that contains the counts of co-occurrences between the $i^{th}$ and $j^{th}$ states of the source and target languages. This language matrix of co-occurrences is the confusion matrix. Figure 1 shows an example of the co-occurrence between state "t_s.en" and state "s_l.ch" when English is taken as the target language. (Note: the Mandarin states and English states are labeled by tag "_ch" and "_en" respectively)

4. Confusion probability calculation: Let $M, N$ be the numbers of states in source and target language. Let $A_{M \times N}$ be the confusion matrix and $A_{ij}$ be the $i^{th}$ row and $j^{th}$ column element of this matrix. Given the target language state $t_j$ and the source language state $s_i$, the confusion probability can be computed as:

$$A_{ij} = \frac{\text{count}(s_i|t_j)}{\sum_{n=1}^{M} \text{count}(s_n|t_j)}$$

$$p_j(\alpha_t) = \alpha p_j(\alpha_{sou}) + (1 - \alpha) \sum_{l=1}^{n} a_{ij} p_l(\alpha_{tar})$$

5. State mapping information: After the confusion probability calculation $A_{ij}$ is obtained, the state mapping information can be derived from this matrix. Given the $i^{th}$ row and $j^{th}$ column element of $A_{ij}$ has the maximal value in the $j^{th}$ column, the $i^{th}$ state from source language is the best matching state to the $j^{th}$ state of the target language. If $k^{th}$ row has the second maximal value in the $j^{th}$ column, the $k^{th}$ state from source language is the 2nd-best matching state to the $j^{th}$ target language state, and so on. Based on this rule, every target language state can find n-best (n< M) matching states from the source language.

6. State-candidate model modification: When the state mapping information and the number of n-best candidates are determined, the new output probability density $p_j(\alpha_t)$ of target language state $j^{th}$ for the observation vector $\alpha_t$ is modified by the n-best state candidates from the source language as follow:

$$p_j(\alpha_t) = \alpha p_j(\alpha_{sou}) + (1 - \alpha) \sum_{l=1}^{n} a_{ij} p_l(\alpha_{tar})$$

Table 3 shows the performances of the nonnative acoustic models with state-candidate model modification approach on the testing database. Model_MM_CI refers to the acoustic models modified by the state-candidate bilingual model modification approach when Model_MAP is regarded as the baseline nonnative English acoustic models and Model_Mand is regarded as the auxiliary acoustic models respectively. "CI", short for "1-Best Candidate", means that each state of baseline acoustic models is modified by just using the best matching state from Mandarin model states via combining the Gaussian mixtures together. As shown, Model_MM_CI performs better than the baseline nonnative English acoustic models, which has a 7.29% relative PhrER reduction.
Table 3: Performances of the nonnative acoustic models with state-candidate model modification approach on the testing database

<table>
<thead>
<tr>
<th>Acoustic Models</th>
<th>PhrER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model_MAP</td>
<td>34.3</td>
</tr>
<tr>
<td>Model_MM_C1</td>
<td>31.8</td>
</tr>
</tbody>
</table>

4.2. N-best candidate selection

Given the state mapping criterion, another aspect concerning model optimality that we have investigated is the number of n-best candidates selected to modify the baseline nonnative acoustic models. Appropriate number of candidates can improve the recognition accuracy further. In our study, different candidate numbers are investigated and compared. Experiment shows that 2-best candidate is the most appropriate choice which can improve the performances significantly.

Table 4: Performances of state-candidate model modification approach with different n-best candidates on the testing database

<table>
<thead>
<tr>
<th>Acoustic Models</th>
<th>PhrER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model_MAP</td>
<td>34.3</td>
</tr>
<tr>
<td>Model_MM_C1</td>
<td>31.8</td>
</tr>
<tr>
<td>Model_MM_C2</td>
<td>31.6</td>
</tr>
<tr>
<td>Model_MM_C3</td>
<td>32.1</td>
</tr>
</tbody>
</table>

Figure 2: Comparison of acoustic modeling performances.

Table 4 presents the experimental results on the testing database when selecting different numbers of n-best candidates. Tags“C1”, “C2” and “C3” refer to three numbers of n-best candidates. As can be seen, Model_MM_C2 outperforms Model_MM_C1 modestly. Even though Model_MM_C3 has most Gaussian components per state, it does not perform best. Model_MM_C2 has the best performance of the three, thus 2-best state candidate is regarded as the preferable number of state candidates to be selected in the state-candidate model modification approach. These experimental results illuminate that most appropriate number of state candidates can give us best performance.

As a whole, nonnative acoustic models whose states are modified with corresponding 2-best state candidates from auxiliary Mandarin models achieve 30.7% and 7.87% relative PhrER reduction respectively, when compared to native English acoustic models. Model_MM_C1, Model_MM_C2 and Model_MM_C3 refer to three numbers of n-best candidates.

5. Conclusions

In the paper we investigated how to improve nonnative speech recognition accuracy with great variations of accented pronunciations, especially when pronunciation substitutions occur. In order to capture characteristics of accents from speakers’ own native language, a novel state-candidate bilingual model modification approach is proposed and presented. Each state of the baseline nonnative acoustic models is modified with 2-best candidate states from auxiliary acoustic models, which are trained by speakers’ mother language. Significant improvement is achieved in our experiments. This approach is proven to be useful for the development of speech recognition systems in which nonnative training data are limited for covering pronunciation variations of testing data.

6. Acknowledgements

This work is partially supported by The National High Technology Research and Development Program of China (863 program, 2006AA010102) MOST (973 program2004CB318106), National Natural Science Foundation of China (10574140, 60535030).

7. References