Speaker Adaptation Method for CALL Systems
Using Bilingual Speakers’ Utterances

Motoyuki Suzuki, Hirokazu Ogasawara, Akinori Ito,
Yuichi Ohkawa* and Shozo Makino.

Graduate School of Engineering, Tohoku University, Japan
*Graduate School of Educational Informatics, Tohoku University, Japan
{moto, ogasawara, aito, kuri, makino}@makino.ecei.tohoku.ac.jp

Abstract
Several CALL systems have two acoustic models to evaluate a learner’s pronunciation. In order to achieve high performance for evaluation, speaker adaptation method is introduced in CALL system. It requires adaptation data of a target language, however, a learner cannot pronounce correctly. In this paper, we proposed two types of new speaker adaptation methods for CALL system. The new methods only require learner’s utterance of the native language for adaptation.

The first method is an algorithm to adapt acoustic models using bilingual’s utterances. The speaker-independent acoustic models of native and target languages are adapted to a bilingual speaker once, then they are adapted to the learner again using the learner’s speech of the native language. Phoneme recognition accuracy is about 5% higher than the baseline method.

The second method is a training algorithm of an acoustic model. It can robustly train bilinguals’ model from a few bilinguals’ utterances. Phoneme recognition accuracy is about 10% higher than the baseline method.

1. Introduction
In recent years, new type of CALL (Computer Assisted Language Learning) systems have been developed (e.g. [1, 2]). These systems can evaluate a learner’s pronunciation using speech recognition technologies.

A learner often makes pronunciation errors which depend on the learner’s native language. For example, a phoneme /v/ in English is pronounced like a similar phoneme /b/ in Japanese by a Japanese native speaker. In order to detect such kind of errors, CALL system has some typical pronunciation error rules, such as /v/ → /b/. When a learner’s utterance is input, various phoneme sequences are made by applying rules to the phoneme sequence of the correct pronunciation, and likelihood is calculated for all phoneme sequences. Finally, the system points out the pronunciation error if the likelihood of the phoneme sequence with pronunciation error is larger than the likelihood of the correct phoneme sequence.

These CALL systems have two acoustic models: the one is for a target language, and the other is for a native language. The evaluation performance depends on the accuracy of the two models. In order to improve speech recognition accuracy, several speaker adaptation methods have been proposed [3, 4]. These methods are also effective for evaluation performance of a learner’s pronunciation, if adaptation data with correct pronunciations can be prepared for both native and target languages. Learner cannot pronounce a target language correctly, although he can pronounce the native language. Therefore, these methods cannot be used for CALL system.

In order to solve this problem, we propose a new speaker adaptation method for CALL systems. It only requires learner’s utterances of the native language as adaptation data.

2. Speaker adaptation method based on bilinguals’ speech data

2.1. Definition of mathematical symbols
Let me define some mathematical symbols here. In this paper, $M^X_Y$ represents an acoustic model for a language $X$ which is trained using speech data uttered by speaker $Y$. For example, $M^L_Y$ represents an acoustic model for a target language, and it is trained using learner’s utterances. $X$ is either $n$ (learner’s native language) or $t$ (target language), and $Y$ is $L$ (a learner), $S_t$ (speaker set for a target language), $S_n$ (speaker set for a native language), $B$ (many bilingual speakers) or $B_i$ (an $i$-th bilingual speaker).

$W^X_{Y→Z}$ represents a regression matrix calculated by MLLR method[4]. It transforms a model $M^X_Y$ into a model $M^X_Z$. The mapping of $M^X_Y$ into $M^X_Z$ using $W^X_{Y→Z}$ is represented by following equations:

$$M^X_Z = W^X_{Y→Z} \cdot M^X_Y \quad (1)$$

2.2. Speaker adaptation method for CALL system
MLLR method[4] is a speaker adaptation methods that can adapt an initial model to the speaker. It requires sev-
eral utterances as adaptation data, and a regression matrix is calculated for several phonemes that appears in the adaptation data (These phonemes are called “known” phonemes in this paper). Finally, “unknown” phoneme models are transformed using the regression matrix.

A speaker-independent model for the native language ($M_{n}^B$) can be adapted using the MLLR framework. However, a speaker-independent model for the target language ($M_{t}^B$) cannot be adapted because a learner cannot utter words/sentences of the target language correctly that are required for the adaptation.

In the previous work[5], we used a regression matrix $W_{S_{n}→L}^n$ for estimating $M_{L}^t$. In this method, all phonemes in the target language are regarded as “unknown” phonemes. From the experimental results, this method was not effective to the estimation of $M_{L}^t$. One of the reason is that the speaker set $S_{t}$ are different from the speaker set $S_{n}$. Figure 1 shows the relationship between models. In this figure, each oval denotes an each model such as $M_{S_{t}}^n$, $M_{L}^t$, and so on. $M_{L}^t$ and $M_{L}^t$ should be almost overlapped because training data of the both models is uttered by the learner. On the other hand, $M_{S_{t}}^n$ and $M_{S_{n}→L}^n$ might not overlap, since a set of the speakers $S_{t}$ is different from $S_{n}$. When the regression matrix $W_{S_{n}→L}^n$ is used to transform the model $M_{S_{n}→L}^n$, obtained model $M_{L}^t$ is far from the learner’s model $M_{L}^t$.

### 2.3. New speaker adaptation method based on bilinguals’ speech data

If we have huge amount of speech data uttered by many bilinguals, native and target model ($M_{B}^n$, $M_{B}^t$) can be trained. In this case, it can be assumed that $W_{B→L}$ is same as $W_{B→L}$ because relative position of $M_{B}^n$ and $M_{B}^t$ is same as that of learner’s models. As the result, the regression matrix $W_{S_{n}→L}^n$ can be used for adapting of the model $M_{B}^t$. However, It is very hard to collect huge amount of bilinguals’ speech data.

In order to solve this problem, we propose a new adaptation method based on a few bilinguals’ speech data. There are two approaches to solve the problem:

1. Estimate the model $M_{St}^t$ from the model $M_{St}^t$ using bilingual’s utterances. $M_{St}^t$ is mapped to $M_{Bt}^t$, once, and then mapped to $M_{St}^t$. MLLR is used as a mapping method.

2. Estimate the model $M_{Bn}^n$ and $M_{Bt}^t$ from a few bilinguals’ speech data. If a speaker-independent model is trained only using a few speakers, the performance of the obtained model is affected by the characteristics of the speakers. In order to normalize these characteristics, SAT[6] is used for training.

In this paper, the new speaker adaptation method based on the first approach is proposed in Section 3, and another method based on the second approach is also proposed in Section 4.

### 3. Speaker adaptation method based on the MLLR mapping

Figure 2 shows the mechanism of the method. A bilingual $B_{t}$ can speak target and native languages correctly. The regression matrices $W_{St→Bt}^t$ and $W_{St→Bt}^t$ can be calculated from utterances in native and target languages respectively. $M_{Bt}^t$ is calculated by applying the $W_{St→Bt}^t$ to the $M_{St}^t$. Then, the model $M_{Bn}^n$ can be estimated by applying the inverse matrix $[W_{St→Bt}^t]^{-1}$ to the model $M_{Bt}^t$. If the relative position between $M_{Bn}^n$ and $M_{Bt}^t$ is same as that of learner’s models, $M_{Bn}^t$ can be obtained by applying the $W_{St→Bt}^t$ to the model $M_{St}^t$.

### 3.1. Details of the method

The speaker adaptation method consists of following 3 steps:

1. Calculate two regression matrices $W_{S_{n}→B_{t}}^n$ and $W_{St→Bt}^t$ for a bilingual. These matrices transform speaker-independent models for native and target languages into bilingual’s models respectively.

2. Calculate a regression matrix $W_{S_{n}→L}$ for learner’s native model.
Table 1: Experimental conditions

<table>
<thead>
<tr>
<th></th>
<th>Japanese database</th>
<th>English database</th>
<th>bilingual data</th>
<th>evaluation data</th>
<th>acoustic model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ATR speech database set C</td>
<td>TIMIT acoustic-phonetic continuous speech corpus</td>
<td>1 male and 2 females</td>
<td>2 males and 2 females</td>
<td>monophone HMM</td>
</tr>
<tr>
<td></td>
<td>32 males and 33 females</td>
<td>326 males and 136 females</td>
<td>30 sentences were uttered</td>
<td>50 Japanese sentences were used</td>
<td>gender dependent model</td>
</tr>
<tr>
<td></td>
<td>100 sentences were uttered by each speaker</td>
<td>10 sentences were uttered by each speaker</td>
<td>for each language</td>
<td>for evaluation</td>
<td>4-state 3-loop, single Gaussian</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3. Estimate the learner’s models $M^n_L$ and $\hat{M}^t_L$ using following equations.

$$M^n_L = W^n_{n \rightarrow L} \cdot M^n_S$$

$$\hat{M}^t_L = \hat{W}^t_{S \rightarrow L} \cdot M^t_S$$

$$\hat{W}^t_{S \rightarrow B_i} = W^n_{n \rightarrow L} \cdot \left[W^n_{S \rightarrow B_i}\right]^{-1} \cdot \hat{W}^t_{S \rightarrow B_i}$$

3.2. Experiments

In order to confirm the effectiveness of the method, speaker adaptation experiments were carried out. The target language was English, and the native language was Japanese. Utterances by three bilinguals (one male and two females) were used for calculation of the regression matrix, and utterances by four Japanese speakers (two males and two females) were used for evaluation. Experimental conditions are shown in Table 1.

Phoneme recognition accuracy was used as evaluation criterion. No linguistic restriction was used in the recognition experiments. In other words, all phonemes of both languages can be followed by any phonemes. Evaluation speech data were labeled by three bilinguals, and pronunciation errors are correctly described in it.

Figure 3 shows the recognition performance. In this figure, “M01” denotes that the male bilingual’s utterances were used for calculation of the regression matrix $W^n_{S \rightarrow B_M01}$ and $W^t_{S \rightarrow B_M01}$, and “F01” and “F02” denotes that the female bilingual speaker was used, and “all” denotes that all speech data uttered by three bilinguals were used for calculation of the regression matrix. “baseline” denotes that only $W^n_{S \rightarrow L}$ was applied to $M^n_S$ and $M^t_S$.

This figure shows that the new speaker adaptation method was more effective than the baseline method. However, “all” was not effective. “all” bilinguals model gave low performance because a speaker characteristics of a bilingual speaker is different from that of other bilingual speakers. (In general, the performance of a speaker-independent model is lower than that of a speaker-dependent model.)

This figure also shows that there were some difference between bilingual speakers. It is caused by the fact that the relative position between a bilingual models, $M^n_{B_i}$ and $M^t_{B_i}$, was different from bilingual to bilingual. The proposed method assumes that the relative position between learner’s models is same as bilingual’s. This result suggests that relative position of “M01” is most similar to that of evaluation speakers.

4. Speaker adaptation method using SAT

4.1. The basic idea

The SAT(Speaker-Adaptive Training) method[6] jointly estimates HMM parameters and regression matrices which adapt the estimated HMM for a training speaker. The resulting model is optimum as long as the model is transformed to a specific speaker using regression matrices obtained by MLLR. Now, if we regard $M^n_{B_i}$ and $M^t_{B_i}$, as one set of HMM and apply SAT method using single regression matrix, the estimated model (that contains both native and target models) will be the best model under assumption that single regression matrix is used for adaptation. It is the basic idea of the proposed method.

The basic scheme of the proposed method is as follows.

1. A set of HMM are created by gathering the
speaker-independent HMMs of native and target languages.

2. The HMM is trained by SAT method using bilinguals’ utterances. In this process, single regression matrix $W_{S_{i}, B_{i}}^{*}$ is used for all distributions.

3. The SAT-trained model is split into native and target HMMs, $\hat{M}_{n B}$ and $\hat{M}_{t B}$.

These HMMs are ready to be adapted by common regression matrix. Therefore, upon adaptation, a regression matrix obtained from native speech $W_{n B \rightarrow L}$ can be applied to both $\hat{M}_{n B}$ and $\hat{M}_{t B}$.

4.2. Experiments

In order to confirm effectiveness of the proposed method, we carried out phoneme recognition experiments. Experimental conditions were same as the previous experiments.

Figure 4 shows the recognition performance. In this figure, “SAT” denotes the result of the proposed method, and “M01” and “baseline” are same as the results in Fig.3. From the experimental results, “SAT” algorithm shows better performance than “baseline” and “M01”. The proposed method can normalize each bilingual model.

5. Conclusions

Several CALL systems have two acoustic models to evaluate a learner’s pronunciation. In order to achieve high performance for evaluation, speaker adaptation method is introduced in CALL system.

Conventional speaker adaptation methods require both native and target speech data with correct pronunciation. However, a learner cannot utter words of the target language in correct pronunciation. In order to solve this problem, we proposed two types of new speaker adaptation methods for CALL system. The new methods only require learner’s utterance of the native language as adaptation data.

The first method is the mapping algorithm of an acoustic model to a target language using bilingual’s utterances. This method adapts the speaker-independent models of native and target languages to a bilingual speaker once. After the mapping, regression matrix calculated by the native speech is used for adaptation of the model. From the experimental results, phoneme recognition accuracy is about 5% higher than the baseline method.

The second method is the training algorithm of an acoustic model. It can robustly acquire the bilingual’s model from a few bilinguals’ utterances. From the experimental results, phoneme recognition accuracy is about 10% higher than the baseline method, and about 5% higher than the first method.

6. References


