ACOUSTIC MODEL ADAPTATION BASED ON PRONUNCIATION VARIABILITY ANALYSIS FOR NON-NATIVE SPEECH RECOGNITION

Yoo Rhee Oh, Jae Sam Yoon, and Hong Kook Kim

Dept. of Information and Communications
Gwangju Institute of Science and Technology (GIST), Gwangju 500-712, Korea
{yroh, jsyoon, hongkook} @ gist.ac.kr

ABSTRACT

In this paper, we investigate the pronunciation variability between native and non-native speakers and propose an acoustic model adaptation method based on the variability analysis in order to improve the performance of a non-native speech recognition system. The proposed acoustic model adaptation is performed in two steps. First, we construct baseline acoustic models from native speech, and perform phone recognition by using the baseline acoustic models to identify most informative variant phonetic units from native to non-native. Next, the acoustic model corresponding to each informative variant phonetic unit is adapted so that the state tying of the acoustic model for non-native speech reflects such a phonetic variability. For further improvement, the traditional acoustic model adaptation such as MLLR or MAP could be applied on the system that is adapted with the proposed method. In this work, we select English as a target language and non-native speakers are all Korean. It is shown from the continuous Korean-English speech recognition experiments that the proposed method can achieve the average word error rate reduction by 12.75% when compared with the speech recognition system with the baseline acoustic models trained by native speech. Moreover, the reduction of 57.12% in the average word error rate is obtained by applying MLLR or MAP adaptation to the adapted acoustic models by the proposed method.

1. INTRODUCTION

Nowadays we have many chances to use a different language from the mother tongue by the stream of the internationalization. Moreover, there is an increasing demand on the automatic systems using the speech recognition. However, the performance of an automatic speech recognition (ASR) system tested by the non-native speech degrades significantly, compared with that by the native speech. The main reason of this problem is that a target language, with which the speech recognition system has been already trained, and the mother tongue of the non-native speaker have different pronunciation spaces of the vowel and consonant sounds. This is because the articulators of the speakers are optimized on their mother tongue by speaking their language repeatedly. Therefore, an ASR for the non-native speech requires kind of adaptation to compensate for this fact.

There have been several research works on non-native speech recognition reported and they can be categorized into one of three approaches: pronunciation modeling, acoustic modeling, and language modeling. In addition, the combination of these approaches can be used for more improvement. The pronunciation modeling makes a non-native speech recognition system to include the pronunciation variants by non-native speakers for each word [1]. On the other hand, the acoustic modeling is usually to adapt the acoustic models by one of adaptation methods such as maximum likelihood linear regression (MLLR), maximum a posteriori (MAP) adaptation [2], and so on. Finally, the language modeling is to improve the ASR performance by adapting the language model [3].

In this paper, we propose an acoustic modeling approach for non-native speech recognition. The main difference from the previous approaches in an acoustic modeling point of view is in that the pronunciation variability is first investigated and then the acoustic model adaptation is performed for the phonetic units that are identified as most variant units from the target language. The pronunciation variability is modeled by a phoneme confusion matrix for pronunciation from native to non-native speech. A phoneme confusion matrix was also introduced in [4] to improve the performance of a non-native ASR system. It, however, was used to merge the acoustic models of non-native language with ones of the target language while we will use the confusion matrix to cluster the state of acoustic models of target language. The proposed acoustic model adaptation method makes the states of the variant units tied. In other words, the proposed method clusters the states with the different central state of the triphones corresponding to each variant phonetic unit. After the proposed adaptation, the mixture of each acoustic model is increased, and for further improvement, the traditional acoustic adaptation method is applied.

The organization of this paper is as follows. In Section 2, the effect of the non-native speech on the native speech baseline ASR is investigated. After that, we propose an acoustic model adaptation method for the improvement of non-native speech recognition in Section 3. Next, the performance of non-native speech recognition using the proposed method is evaluated and compared with those using traditional acoustic model adaptation methods in Section 4. Finally, we conclude our findings in Section 5.

2. EFFECT OF NON-NATIVE SPEECH

The performance of the ASR system tested by the non-native speakers tends to be degraded markedly since the non-native speakers make the pronunciation variants. In this section, the effect of the non-native speech on the performance of the ASR system constructed from native speech is discussed. To do this, we first construct an English baseline ASR system and then evaluate the ASR performance of the English baseline using the English spoken by Koreans.

1-4244-0469-X/06/$20.00 ©2006 IEEE

2.1. English baseline ASR

A subset of the Wall Street Journal database [5], WSJ0, is used as a training set for the native-English ASR system. WSJ0 is a 5000-word closed-loop task to evaluate the performance of a large vocabulary continuous speech recognition (LVCSR) system. The training set consists of 7,138 utterances recorded by the Sennheiser close talking microphone and several far talking microphones, where all the utterances are sampled at a rate of 16 kHz.

As a recognition feature, we extract 12 mel-frequency cepstral coefficients (MFCC) with a logarithmic energy for every 10 ms analysis frame, and concatenate their first and second derivatives to obtain a 39-dimensional feature vector. During the training and testing, we apply cepstral mean normalization and energy normalization to the feature vectors.

The acoustic models are based on the 3-state left-to-right, context-dependent, 4-mixture, and cross-word triphone models, and trained by using the HTK version 3.2 toolkit [6]. All the triphone models are expanded from 41 monophones including a silence and a pause model and the states of triphone models are tied by employing a decision tree [7]. As a result, we have 8,360 triphones and 5,356 states, which is referred to as AM0.

2.2. Speech database for English spoken by Koreans

In this work, we use a subset of the Korean-Spoken English Corpus (K-SEC) [8], which is composed of the English pronunciations spoken by Korean and native speakers. This database is divided into three parts: one is used for developing the proposed method that will be described in Section 3, and the others are used for the evaluation of the performance of English baseline ASR and the proposed method, respectively. In other words, two evaluation sets are composed of utterances spoken by 49 Koreans and 7 native speakers, respectively.

Utterances from 7 Koreans are used for the development set, where each Korean speaker pronounces 435 isolated words and 36 sentences whose average number of words is about 5.4. Thus, we have 3,045 isolated words and 252 continuous sentences. On the other hand, the two evaluation sets are made up with continuous sentences, where each Korean or native speaker utters 10 continuous sentences, which results in the total number of 86 sentences. In other words, we have 490 utterances and 70 utterances for non-native speech and native speech, respectively.

2.3. Effect of native and non-native speech on the baseline ASR

In order to explore a behavior of acoustic models by the difference between the target language and the mother tongue, we take the lexicon only from the text of the test set. The pronunciation of each word is built from the CMU pronunciation dictionary [9] and the missing words in the CMU dictionary are transcribed manually. A backed-off bigram is used for a language model.

The performance of the baseline ASR described in Section 2.1 is tested by the two evaluation sets. It is shown that the average word error rate (WER) of the English baseline ASR system is 4.21% and 39.22%, when the ASR system is tested by native speakers and by non-native speakers, respectively. This result verifies the fact that the performance of the ASR system tested by the non-native speech could be degraded exceedingly.

3. ACOUSTIC MODEL ADAPTATION FOR NON-NATIVE SPEECH RECOGNITION

In this section, an acoustic model adaptation method is proposed to improve the recognition performance of the baseline ASR system tested by the non-native speech. The proposed method consists of two steps: the analysis of the pronunciation variability of the non-native speech from the native speech using a phoneme confusion matrix, and the acoustic model adaptation based on the analysis of the pronunciation variants. Fig. 1 shows the overall procedure of the proposed acoustic model adaptation method. The left part of Fig. 1 shows the procedure of constructing the English baseline ASR system and analyzing the pronunciation variants, which is describe in Section 3.1. On the other hand, the right part of Fig. 1 shows the procedure of the proposed acoustic model adaptation method, which is also described in Section 3.2.

3.1. Analysis of the pronunciation variability

The pronunciation variability of the non-native speakers can be investigated on the basis of a broad knowledge about the target language and the mother tongue of the non-native speakers. This approach is generally acceptable but it has difficulties in dealing with the real pronunciation effects because, for example, some of Koreans can pronounce English as if they are native. Instead of using such a knowledge-based approach, we consider a data-driven approach to analyze the pronunciation variability.

The first step is to recognize the non-native speech on the English baseline ASR system (AM0) and to obtain the relationship between the target pronunciation and the incorrectly recognized pronunciation. It is assumed here that the most of incorrectly recognized pronunciations correspond to the pronunciation variants of the non-native speakers. The second step is to generate a phoneme confusion matrix of the recognition result. In the phoneme confusion matrix, each element has the value of a relative frequency of the incorrectly recognized pronunciation for a given target

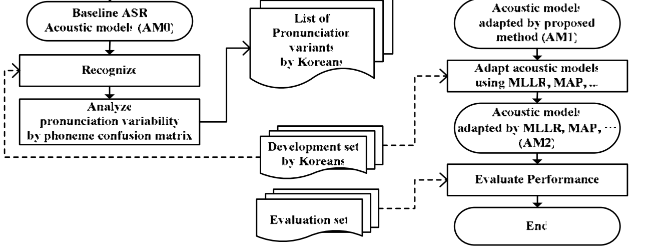
Start
Generate monophone acoustic models
Native English training set
Generate triphone acoustic models
Intermediate Acoustic models
Cluster states of acoustic models
Increase mixture of acoustic models
Baseline ASR Acoustic models (AM0)
Recognize
Analyze pronunciation variability by phoneme confusion matrix
List of Pronunciation variants by Koreans
Development set by Koreans
Evaluation set
Cluster states of acoustic models
Increase mixture of acoustic models
Acoustic models adapted by proposed method (AM1)
Adapt acoustic models using MLLR, MAP, ...
Acoustic models adapted by MLLR, MAP, ··· (AM2)
Evaluate Performance
End

Figure 1: The overall procedure of the proposed adaptation method where the left part is for the analysis of the pronunciation variability using the baseline ASR system and the right part is the proposed acoustic model adaptation method based in the state-tying step.

pronunciation. In other words, the entry $a_{p(i),p(j)}$ is calculated by counting occurrence that phone $p(i)$ is recognized as the phone $p(j)$ and normalizing it by the total number of the phone $p(i)$ that is labeled in the text of the development set. Therefore, the value of an off-diagonal element corresponds to the degree of pronunciation variant for each target pronunciation. From this observation, we select informative variants among the pronunciation variants if their value in the confusion matrix is greater than a threshold. In this work, we empirically set the threshold as 0.16 and 0.13 when the target phoneme is a consonant and a vowel, respectively. Actually, the comparative study on Korean and English phonetics strongly supports our selection rule. In other words, the pronunciation variant obtained from the phoneme confusion matrix has a strong correlation with that from the comparative study on Korean and English phonetics [10].

3.2. Proposed acoustic model adaptation for non-native speech recognition

As shown in Fig. 1, the proposed acoustic model adaptation method for non-native speech recognition is performed with the intermediate acoustic models that are basically triphone models with a single mixture. In general, the intermediate acoustic models are clustered using a decision tree and each clustered group on the leaf node in the tree is tied with the representative, which has broadest variances among the clustered models. In the proposed adaptation method, however, each phone is differently dealt with in the acoustic model clustering stage whether or not it has a pronunciation variant from the pronunciation variability analysis described in Section 3.1. For example, there is a pronunciation variability denoted by /a/→/b/, which means that a phonetic unit /a/ is mostly mis-recognized as /b/ so that /a/ is phonetically varied into /b/ for non-native speech. In this case, the acoustic models for both /a/ and /b/ are pooled together on the root node of the decision tree for the phone /a/. However, for a phone (/c/) which has no pronunciation variant, the acoustic models of the triphones including /c/ as a central phone are pooled on the root node of the decision tree for the phone /c/.

Fig. 2 illustrates how the proposed acoustic model adaptation method works in a view of the state-tying step. Fig. 2(a) shows a decision tree for the phone /P/ that has no pronunciation variants. In this case, the acoustic models for only /P/ are pooled on the root node of the decision tree. Fig. 2(b) shows a decision tree for the phone /IY/ that has a pronunciation variant /IH/ (/IY/→ /IH/). Therefore, the acoustic models of the triphones including both /IY/ and /IH/ as central phones are pooled on the root node of the decision tree. After clustering all the acoustic models using the decision tree, the clustered acoustic models in each leaf node of the decision tree are tied with the representatives.

Finally, we increase the number of mixtures for the adapted acoustic models. Especially, we apply MLLR or MAP to the acoustic models for the further improvement of the performance of non-native speech recognition.

4. EXPERIMENTS

In this section, we evaluate the performance of the proposed acoustic model adaptation method for a non-native ASR system and compare it with several acoustic model adaptation methods. For the comparison, the English baseline ASR system is adapted with the development set by retraining (AM0-Retrain), and then further adapted by one of the traditional acoustic adaptation methods (MLLR or MAP) or their combination.

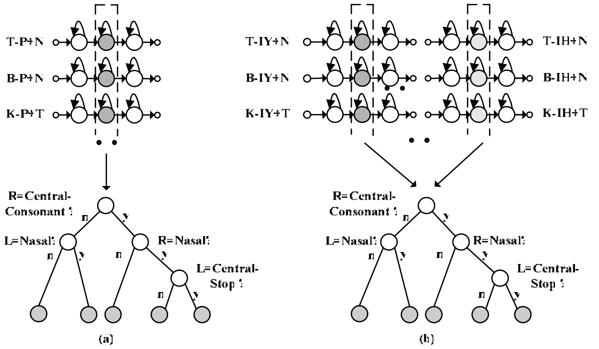

Figure 2: An example of a decision tree for state-tying acoustic models throughout the proposed acoustic model adaptation method. (a) the phone /P/ without any pronunciation variants and (b) the phone /IY/ with a pronunciation variant /IH/.

First, to apply the proposed acoustic model adaptation method, we analyzed the pronunciation variants of the Korean-English by recognizing the development set with the English baseline ASR system (AM0). As a result, the six informative pronunciation variants were obtained from the confusion matrix. They were /CH/→/T/, /IH/→/IY/, /OY/→/IY/, /ER/→/R/, /UH/→/OW/, and /EH/→/AE/. Second, the proposed acoustic model adaptation method was applied to the intermediate acoustic models in the state-tying step using these six informative pronunciation variants, which resulted in the acoustic models (AM1).

Table 1 shows the performance comparison of the baseline acoustic models (AM0), the retrained acoustic models (AM0-Retrain), and the adapted acoustic models by the proposed acoustic model adaptation method (AM1). All the performances were measured by the average word error rate (WER). By comparing the first and second rows in the table, it is shown that retraining acoustic models significantly degraded the ASR performance by native speakers even it improved the ASR performance by non-native or Korean speakers. On the average, the WER was increased by 58.77%. On the other hands, the proposed acoustic model adaptation method achieved the average WER reduction by 12.75% compared with AM0, while it gave little degradation of the performance for the native English speech. It is here concluded that the proposed method could improve the ASR performance for

Table 1: Comparison of word error rates (%) of the baseline ASR system (AM0), the retrained version of the baseline ASR system with the development set by Koreans, and the acoustic models (AM1) by applying the proposed acoustic model adaptation method.

Evaluation Set Acoustic Models		Korean	Native	Avg.
Baseline ASR system (AM0)		39.22	4.21	21.71
Retrained baseline ASR system with development set (AM0-Retrain)		26.87	42.07	34.47
Proposed adaptation (AM1)	Adapted based on the variability such as /CH/→/T/,/IH/→/IY/, /OY/→/IY/,/ER/→/R/, /UH/→/OW/,/EH/→/AE/	33.18	4.7	18.94

the non-native speech while it maintained the ASR performance for the native speech.

Next, we applied the traditional acoustic model adaptation methods to AM1 in order to improve the ASR performance for the non-native speech. In addition, we applied the same technique to AM0 for a fair performance comparison. The methods applied in this paper were MLLR, MAP, and the second pass adaptation with the combination of MLLR and MAP. Table 2 shows the performance comparison of the ASR system employing six different acoustic models. We first applied MAP adaptation to AM0 and AM1 showed the WERs of the ASR system using AM0+MAP and AM1+MAP into the first and second rows of Table 2, respectively. Compared to the result in Table 1, MAP could achieve the average WER reduction by 54.86% and 55.18% with little degradation for the native speech when it was applied to AM0 and AM1, respectively. On one hand, MLLR could reduce the average WERs by 40.99% and 52.56% after adapting AM0 and AM1, respectively. In this work, MAP provided a little better adaptation performance than MLLR, while the difference was marginal.

Finally, we applied the second pass adaptation of the combination of MLLR and MAP to AM0 and AM1. The WERs of the ASR system using AM0+MLLR+MAP and AM1+MLLR+ MAP are shown in the last two rows of Table 2, respectively. In other words, MLLR was first applied to AM0 or AM1, and then MAP was performed. Compared to the performance of AM0 and AM1 as shown in Table 1, MLLR+MAP reduced the average WERs by 52.74% and 57.12% when it was applied to AM0 and AM1, respectively. It was shown from the table that the second pass adaptation could reduce WER for non-native speech than MLLR or MAP adaptation only. Moreover, AM1+MLLR+MAP gave the lowest WER among the six acoustic models.

From Table 1 and Table 2, it could be concluded as follows. First, the proposed acoustic model adaptation method improves the performance. That is, it reduces not only the WER for the non-native speech, but also the degradation for the native speech. Second, the traditional adaptation provides the more powerful performance on the ASR system for both the native and non-native speech, especially the second pass adaptation with the combination of MLLR and MAP.

5. CONCLUSION

In this paper, we proposed the acoustic model adaptation method for non-native speech recognition, especially English speech recognition spoken by Korean. The proposed method, which is a data-driven approach, first ranked the phonetic units that gave most informative pronunciation variability by recognizing non-native speech using the acoustic models trained by native speech. And then, the states of the acoustic models for the phonetic units with the informative pronunciation variants were differently tied from those without the pronunciation variants. From the continuous Korean-English speech recognition experiments, it was shown that the proposed acoustic mode adaptation method achieved the average WER reduction by 12.75% compared to the English baseline ASR system. In addition, the proposed acoustic model adaptation method maintained the performance of the ASR system tested by the native speakers. In order to achieve further performance improvement of ASR, the traditional acoustic model adaptation methods such as MLLR and MAP were applied to the acoustic models that were already adapted by the proposed acoustic model adaptation method. As a result, it was shown that MLLR followed by MAP adaptation could provide the best performance and reduced the average WER by 57.12% compared to the baseline ASR system.

Table 2: Comparison of word error rates (%) of the ASR system employing the acoustic models adapted by MLLR or MAP on AM0 (baseline acoustic models) and AM1 (acoustic models adapted by the proposed method).

Evaluation Set Acoustic Model	Korean	Native	Avg.
Baseline (AM0) + MAP	14.71	4.89	9.80
Proposed adaptation (AM1) + MAP	13.89	5.57	9.73
Baseline (AM0) + MLLR	18.77	6.85	12.81
Proposed adaptation (AM1) + MLLR	15.02	5.57	10.30
Baseline (AM0) + MLLR + MAP	13.07	7.44	10.26
Proposed adaptation (AM1) + MLLR + MAP	11.78	6.84	9.31

6. ACKNOWLEDGEMENTS

This work was supported by Korea Research Foundation Grant (KRF-2005-202-D00367), by the Ministry of Information and Communication (MIC) through the Realistic Broadcasting Research Center (RBRC) at Gwangju Institute of Science and Technology (GIST), and by the Korean Science and Engineering Foundation (KOSEF) through the Immersive Contents Research Center (ICRC) at Gwangju Institute of Science and Technology (GIST).

7. REFERENCES

[1] N. Binder, R. Gruhn, and S. Nakamura, "Recognition of non-native speech using dynamic phoneme lattice processing," in *Proc. Spring meeting of the Acoustical Society of Japan*, pp. 203–204, Mar. 2002.

[2] G. Zavagliakos, R. Schwartz, and J. McDonough, "Maximum a posteriori adaptation for large scale HMM recognizers," in *Proc. ICASSP*, Atlanta, GA, pp. 725–728, May 1996.

[3] J. Bellegarda, "An overview of statistical language model adaptation," in *Proc. ISCA Workshop on Adaptation Methods for Speech Recognition*, Sophia-Antipolis, France, pp. 165–174, Aug. 2001.

[4] J. Morgan, "Making a speech recognizer tolerate non-native speech through Gaussian mixture merging," in *Proc. InSTIL/ICALL Symposium on Computer-Assisted Language Learning*, Venice, Italy, pp. 213–216, June 2004.

[5] D. Paul and J. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proc. DARPA Speech and Language Workshop*, Arden House, NY, pp. 357-362, Feb. 1992.

[6] S. Young, *et al*, "The HTK Book (for HTK Version 3.2)," Microsoft Corporation, Cambridge University Engineering Department, Dec. 2002.

[7] S. Young, J. Odell, and P. Woodland, "Tree-based state tying for high accuracy acoustic modeling," in *Proc. ARPA Human Language Technology Workshop*, Princeton, NJ, pp. 307-312, Mar. 1994.

[8] S.-C. Rhee, S.-H. Lee, S.-K. Kang, and Y.-J. Lee, "Design and construction of Korean-spoken English corpus (K-SEC)," in *Proc. ICSLP*, Jeju Island, Korea, pp. 2769-2772, Oct. 2004.

[9] H. Weide, "The CMU Pronunciation Dictionary, release 0.6," Carnegie Mellon University, 1998.

[10] H.-M. Youe, "A survey of the Korean learners' problems in mastering English pronunciation," in *Malsori (Journal of the Phonetic Society of Korea)*, vol. 42, pp. 47-56, Jun. 2001.