NON-NATIVE PRONUNCIATION VARIATION MODELING USING AN INDIRECT DATA DRIVEN METHOD

Mina Kim, Yoo Rhee Oh, and Hong Kook Kim

Department of Information and Communications
Gwangju Institute of Science and Technology (GIST), Gwangju 500-712, Korea
{kma58, yroh, hongkook} @ gist.ac.kr

ABSTRACT

In this paper, we propose a pronunciation variation modeling method for improving the performance of a non-native automatic speech recognition (ASR) system that does not degrade the performance of a native ASR system. The proposed method is based on an indirect data-driven approach, where pronunciation variability is investigated from the training speech data, and variant rules are subsequently derived and applied to compensate for variability in the ASR pronunciation dictionary. To this end, native utterances are first recognized by using a phoneme recognizer, and then the variant phoneme patterns of native speech are obtained by aligning the recognized and reference phonetic sequences. The reference sequences are transcribed by using each of canonical, knowledge-based, and hand-labeled methods. Similar to non-native speech, the variant phoneme patterns of non-native speech can also be obtained by recognizing non-native utterances and comparing the recognized phoneme sequences and reference phonetic transcriptions. Finally, variant rules are derived from native and non-native variant phoneme patterns using decision trees and applied to the adaptation of a dictionary for non-native and native ASR systems. In this paper, Korean spoken by Chinese native speakers is considered as the non-native speech. It is shown from non-native ASR experiments that an ASR system using the dictionary constructed by the proposed pronunciation variation modeling method can relatively reduce the average word error rate (WER) by 18.5% when compared to the baseline ASR system using a canonical transcribed dictionary. In addition, the WER of a native ASR system using the proposed dictionary is also relatively reduced by 1.1%, as compared to the baseline native ASR system with a canonical constructed dictionary.

***Index Terms*—** Speech recognition, pronunciation variation, non-native speech recognition, indirect data-driven approach, dictionary adaptation

1. INTRODUCTION

With increasing globalization, the need for effective inter-lingual communication has also grown. However, most people speak foreign languages with variant or influent pronunciation, thereby leading to an increased demand for the development of non-native speech recognition systems [1]. However, research has shown that the performance of non-native automatic speech recognition (ASR) systems significantly degrades, as compared with the performance of native ASR systems. To this end, there are three major approaches for handling non-native speech for ASR: acoustic model adaptation, pronunciation model adaptation, and language model adaptation [2]. In this paper, we focus on pronunciation modeling as a means of improving the performance of non-native ASR systems.

There have been a multitude of proposals associated with model pronunciation variations for non-native speech. They can generally be divided into two categories: knowledge-based approaches and data-driven approaches [3]. As representatives of knowledge-based approaches, Downey *et al.* [4] and Tajchman *et al.* [5] generated pronunciation rules from phonological knowledge to develop a pronunciation dictionary based on pronunciation rules. However, a notable drawback of their approaches was that the rules were often very general, resulting in too many variants in the dictionary. Their use of these rules was also quite time-consuming and did not cover all aspects of non-native speech.

For this reason, it can be said that a data-driven approach is a more preferable method for modeling pronunciation variations [3] because this method attempts to derive pronunciation variants directly from speech signals. In [6], variants were derived using a phoneme recognizer and pronunciation rules were constructed using a decision tree. Confidence measures were then used to select only the most reliable variants from among all the recognized variants; a similar approach was applied in the Vermobil project by Wolff *et al*. [7]. In addition, Amdal *et al.* [8] examined non-native speech using a phoneme recognizer to determine variants, and removed variants caused by recognition errors based on statistics pertaining to the co-occurrences of phonemes. Goronzy *et al.* [9] also used an English phoneme recognizer to generate English pronunciations for German words and used decision trees that were able to predict English-accented variants from German canonical transcriptions.

In such traditional data-driven methods described above, the pronunciation variants obtained were dependent on pronunciation training databases [3][6][8]. Therefore, even though the performance of non-native ASR improved through the application of a traditional data-driven approach, the recognition accuracy of native ASR could actually be reduced because of the increased confusability in native pronunciation. Thus, the goal of this paper is to improve the performance of non-native ASR while maintaining the performance of native ASR.

A sub-division into direct and indirect data-driven methods can be applied. Directly modeling the pronunciation of each vocabulary word from training data requires all vocabulary words to be well represented in the training data. On the other hand, an indirect data-driven approach uses the training data to derive pronunciation rules that in turn can be applied to generate one or

978-1-4244-1746-9/07/$25.00 ©2007 IEEE

more baseforms of any vocabulary word [10]. In this paper, we propose a new pronunciation modeling method based on an indirect data-driven method, where pronunciation variability is investigated from the training speech data, and variant rules are then derived and applied to compensate for variability in the ASR pronunciation dictionary. Specifically, pronunciation variant rules are obtained from both native and non-native speech databases. First, native utterances are recognized by using a phoneme recognizer, and then variant phoneme patterns of native speech are obtained by aligning the recognized phonetic sequences and their corresponding reference phonetic sequences. The reference sequences can then be transcribed by one of three methods such as the canonical, knowledge-based, or hand-labeled method. Similar to non-native speech, variant phoneme patterns of non-native speech are also obtained by recognizing non-native utterances and aligning the recognized phoneme sequence and reference phonetic transcriptions. Finally, variant rules are derived from the native and non-native variant phoneme patterns using a decision tree and subsequently applied to the adaptation of a dictionary used for non-native and native ASR systems.

Following this Introduction, Section 2 gives an overview of the speech databases and the baseline ASR system used in this paper. Section 3 proposes a pronunciation variation modeling method, and Section 4 uses an example to describe how the proposed pronunciation variation modeling method works. Section 5 presents a performance evaluation and comparison of the native and non-native ASR systems with the baseline system, based on the proposed method. Finally, we conclude our findings in Section 6.

2. SPEECH DATABASE AND BASELINE ASR SYSTEM

2.1. Speech database

This subsection describes the two databases used in this paper: a native speech database, and a non-native speech database [1]. Basically, Korean is the native spoken language, and Korean spoken by Chinese speakers is the non-native one.

A large vocabulary continuous Korean speech database, referred to as CleanSent01 [11], is used for training a native ASR system, developing the proposed pronunciation modeling method, and evaluating the baseline performance of the native ASR system. The CleanSent01 database consists of 20,806 sentences spoken by 200 Koreans; 100 males and 100 females. The database is divided into three sets: a training set, a development set, and a test set. The training set used to train the acoustic models is composed of all the utterances of 170 speakers, resulting in 17,996 utterances with 30,633 different words. The development set is used to develop the proposed method, consisting of 2,132 utterances with 4,159 different words. The remaining 200 utterances are used for the evaluation of the baseline performance of the native ASR system.

For non-native ASR, a subset of the foreign-spoken Korean database, referred to as F-Korean01 [12], is used for developing the proposed method and evaluating the performance of the non-native ASR. The F-Korean01 database is composed of 2,979 utterances spoken by 10 male and 10 female native Chinese speakers. The development set is composed of 1,479 utterances from 10 speakers, and the remaining 1,500 utterances from 10 speakers are used as the test set.

Table 1: List of Korean phonemes for native and non-native ASR.

Vowel (21) (Jungseong)	Monophthong (9)	ㅣ(i), ㅔ(e), ㅐ(E), ㅜ(u), ㅗ(o), ㅏ(a), ㅡ(U), ㅓ(v), ㅚ(O)
	Diphthong (12)	ㅟ(wi), ㅞ(we), ㅙ(jE), ㅝ(wv), ㅘ(wa), ㅖ(je), ㅒ(jE), ㅑ(ja), ㅕ(jv), ㅛ(jo), ㅠ(ju), ㅢ(xi)
Consonant (19) (Choseong, Jongseong)		ㄱ(g), ㄲ(G), ㄴ(n), ㄷ(d), ㄸ(D), ㄹ(l), ㅁ(m), ㅂ(b), ㅃ(B), ㅅ(s), ㅆ(S), ㅈ(z), ㅉ(Z), ㅊ(c), ㅋ(k), ㅌ(t), ㅍ(p), ㅎ(h), ㅇ(N - jongseong)

Table 2: Comparison of the average word error rates (%) of the baseline ASR system using the dictionaries obtained by canonical (CC_Dict), knowledge-based (KB_Dict), and hand-labeled (HL_Dict) transcriptions.

Dictionary	Non-native (F-Korean01)	Native (CleanSent01)
CC_Dict	28.33	43.47
KB_Dict	27.73	34.43
HL_Dict	27.73	35.00

2.2. Baseline ASR system

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

The acoustic models are based on the 3-state left-to-right, context-dependent, 4-mixture, and cross-word triphone models, and trained using the HTK Version 3.2 Toolkit [13]. All the triphone models are expanded from 42 monophones, which include silence and a short pause model, and the triphone models states are tied by employing a decision tree [14]. As a result, we obtain 10,138 triphones and 11,807 states.

Table 1 shows the 40 phonemes used for the Korean ASR system except for silence and short pauses. It is noticed in the table that each syllable in Korean can be divided into one of three components: Choseong, Jungseong, or Jongseong.

2.3. Performance evaluation of the baseline ASR system

We evaluated the performance of the baseline native and non-native ASR systems by using the test set of the CleanSent01 database for native speech and the test set of the F-Korean01 database for non-native speech, respectively. Here, three respective baseline dictionaries for ASR were obtained by canonical, knowledge-based, and hand-labeled transcriptions, resulting in 566,260 entries for native speech and 353 entries for non-native speech. Typically, the knowledge-based transcription generated phonetic sequences from the Romanization form using

[1] All the utterances from native and non-native speech databases were recorded under clean conditions.

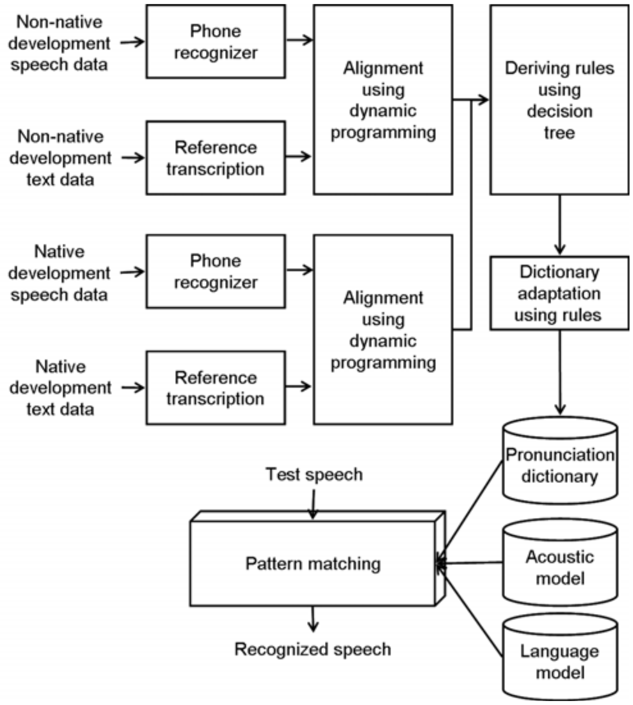

Non-native development speech data
Phone recognizer
Non-native development text data
Reference transcription
Alignment using dynamic programming
Deriving rules using decision tree
Native development speech data
Phone recognizer
Native development text data
Reference transcription
Alignment using dynamic programming
Dictionary adaptation using rules
Pronunciation dictionary
Test speech
Pattern matching
Acoustic model
Recognized speech
Language model

Figure 1**:** Procedure for the proposed pronunciation variation modeling method based on an indirect data-driven approach applied to native and non-native speech.

phonological rules provided by a Korean pronunciation rulebook [15]. Here, the total number of phonological rules was 90, selected from the Korean pronunciation rulebook. The hand-labeled transcription was provided by experts.

Table 2 shows the average word error rates (WERs) of ASR systems using the three dictionaries. For the canonical dictionary denoted as CC_Dict, the WERs of the ASR system for non-native speech (F-Korean01) and native speech (CleanSent01) were 28.33% and 43.47%, respectively. The WER of native speech was higher than that of non-native speech due to the higher complexity of CleanSent01 than F-Korean01 since the vocabulary size of CleanSent01 was significantly larger than that of F-Korean01. However, the WERs of the ASR system using the knowledge-based dictionary (KB_Dict) and the hand-labeled dictionary (HL_Dict) were reduced to 34.43% and 35.00%, though the WERs for non-native speech remained little changed. The reason why KB_Dict and HL_Dict had lower WERs than CC_Dict was that KB_Dict and HL_Dict incorporated some degree of pronunciation variation for native speech from the phonological knowledge and the experiences of speech experts, respectively. In the proposed pronunciation variation modeling method, KB_Dict and HL_Dict will be used to transcribe reference sequences of native and non-native development databases to improve the ASR performance of native speech, while CC_Dict will be only used to compare ASR performance.

3. PRONUNCIATION ADAPTATION FOR NON-NATIVE SPEECH

An indirect data-driven approach for non-native speech, proposed in [16], used training data to adapt a dictionary for non-native pronunciation variation from rules obtained by forced alignment or phoneme recognition. However, variant rules for non-native speech were only utilized for the dictionary adaptation of non-native ASR. However, non-native pronunciation is different from native pronunciation due to the different intonations, phonological processes, and pronunciation rules inherent in the speaker's mother tongue [17]. Thus, if the adapted dictionary for non-native speech is used for native ASR, the performance of the native ASR system could degrade due to an increase in confusability [18].

In order to mitigate this problem, we propose a pronunciation dictionary adaptation method that combines the variant rules obtained from native and non-native speech to improve the performance of non-native ASR, while maintaining native ASR performance. Fig. 1 shows the procedure for the proposed pronunciation variation modeling method based on an indirect data-driven approach applied to native and non-native speech. The five steps of the procedure are as follows.

Step 1) Each utterance in the development set of native speech is recognized by using a phoneme recognizer. The recognized phoneme sequence is aligned using a dynamic programming algorithm with one of the reference phoneme sequences of the utterance obtained by the canonical (CC_Dict), knowledge-based (KB_Dict), and hand-labeled (HL_Dict) transcriptions.

Step 2) Step 1) is repeated using all the utterances in the non-native development set.

Step 3) By using the alignment results of Steps 1) and 2), variant phoneme patterns are obtained.

Step 4) Pronunciation variation rules are derived from the variant phoneme patterns using a decision tree.

Step 5) Pronunciation variations are generated from the pronunciation variation rules, and a dictionary is adapted for both native and non-native ASR.

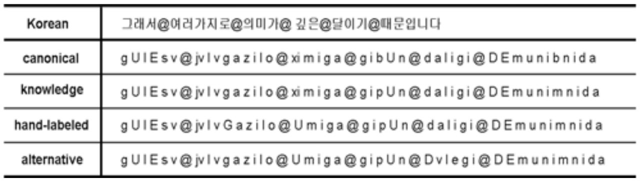
The following two subsections provide further details of the steps of the procedure described above.

3.1. Phoneme recognition and alignment sequence

To derive the pronunciation rules, we first construct a phoneme recognizer by using the acoustic models described in Section 2.2. The acoustic model for the phoneme recognizer includes 10,138 triphones, and a back-off bigram language model is used for recognition, where a list of 42 phonemes with silence and short pauses is used as the dictionary for phoneme recognition. After that, all the utterances in the development set of the CleanSent01 and F-Korean01 databases are recognized using a phoneme recognizer. In this way, we obtain a list of phoneme sequences. However, there are no word boundaries in the list, which are required to differentiate inter-word pronunciation variations from cross–word pronunciation variations [19]. To obtain these word boundaries, the recognized phoneme sequence is aligned on the basis of a dynamic programming algorithm and compared with one of the reference transcriptions with word boundaries. Throughout the remainder of this paper, @ indicates a word boundary.

From the alignment between the recognized phoneme sequence and a reference transcription, a rule pattern is obtained if the following condition is satisfied:

Table 3: Example of three reference sequences obtained by canonical, knowledge-based, and hand-labeled transcriptions, and an alternative phonetic sequence after recognizing a Korean utterance: “그래서 여러가지로 의미가 깊은 달이기 때문입니다,” which in English means “This is because this month has several deep meanings.”



Korean	그래서@여러가지로@의미가@ 깊은@달이기@때문입니다
canonical	gUlEsv@jvlvgazilo@ximiga@gibUn@daligi@DEmunibnida
knowledge	gUlEsv@jvlvgazilo@ximiga@gipUn@daligi@DEmunimnida
hand-labeled	gUlEsv@jvlvGazilo@Umiga@gipUn@daligi@DEmunimnida
alternative	gUlEsv@jvlvgazilo@Umiga@gipUn@Dvlegi@DEmunimnida

Table 4: The rule pattern is obtained using Eq. (1) for the sentence in Table 3.

@-@-g+U+l→g	i-l-o+@+@→o	a-l-i+g+i→e
@-g-U+l+E→U	@-@-xi+m+i→U	l-i-g+i+@→g
g-U-l+E+s→l	@-xi-m+i+g→m	i-g-i+@+@→i
U-l-E+s+v→E	xi-m-i+g+a→i	@-@-D+E+m→D
l-E-s+v+@→s	m-i-g+a+@→g	@-D-E+m+u→E
E-s-v+@+@→v	i-g-a+@+@→a	D-E-m+u+n→m
@-@-jv+l+v→jv	@-@-g+i+b→g	E-m-u+n+i→u
@-jv-l+v+g→l	@-g-i+b+U→i	m-u-n+i+b→n
jv-l-v+g+a→v	g-i-b+U+n→p	u-n-i+b+n→i
l-v-g+a+z→g	i-b-U+n+@→U	n-i-b+n+i→m
v-g-a+z+i→a	b-U-n+@+@→n	i-b-n+i+d→n
g-a-z+i+l→z	@-@-d+a+l→D	b-n-i+d+a→i
a-z-i+l+o→i	@-d-a+l+i→v	n-i-d+a+@→d
z-i-l+o+@→l	d-a-l+i+g→l	i-d-a+@+@→a

$$L_1 - L_2 - X + R_1 + R_2 \to Y \qquad (1)$$

where X is a phoneme that is to be mapped into Y, and the left and right phonemes in the reference transcription are L_1 and L_2, and R_1 and R_2, respectively.

As is known, it is rather difficult to differentiate pronunciation variations from the substitution, deletion, and insertion errors incurred during phoneme recognition [9]. Therefore, recognition errors need to be as small as possible. In this paper, these errors are reduced in two ways. First, we perform a Viterbi search by using 100-best lists, which improves the performance of phoneme recognition from 71.5% to 76.8%. Second, if more than half of the neighboring phonemes of X in Eq. (1) are different from the neighboring phonemes of the target phoneme Y, the rule pattern is removed from the rule pattern set.

3.2. Deriving rules using a decision tree and adapting a dictionary

Decision tree modeling is a popular method for deriving pronunciation variation rules [6][7]. In this paper, we use C4.5, a software extension of the basic ID3 algorithm designed by Quinlan [20]. After the rule formulation is categorized by filtering errors, pronunciation variation rules are constructed by C4.5. Their attributes are the two left phonemes, L_1 and L_2, and the two right phonemes, R_1 and R_2 [from Eq. (1)], of the affected phoneme X. The output class is the target phonemes, where one decision tree is constructed for each phoneme. After the decision tree is built based on the established rule formulations and filtering the phoneme-to-phoneme mapping between the two transcriptions, we then construct rule sets for each phoneme using options provided by C4.5. Consequently, we obtain 376 rules from the decision tree training.

From the rules obtained using the decision tree, our proposed dictionary can thus be derived. However, the addition of pronunciation variants to a dictionary increases the confusability, especially if the dictionary is large. This large increase in confusability is probably a cause of only small improvements or even deteriorations of ASR performance; by appropriately selecting pronunciation variations, this confusability can be reduced. Fortunately, C4.5 provides an accuracy of each rule, which is used for the selection of rules in this paper. In other words, we select a rule if the accuracy of the rule is greater than 0.8 for the non-native development database, and 0.6 for the native development database, resulting in a total of 263 rules.

4. EXAMPLE OF PRONUNCIATION MODELING

This section shows how the proposed pronunciation variation modeling method works by presenting a detailed example according to the five steps described in Section 3. Specifically, Section 4.1 presents the first three steps, and Section 4.2 shows the remaining two steps.

4.1. Phoneme recognition and alignment sequence for native and non-native speech

As a first step of the proposed method, the reference transcriptions are obtained in three different ways: canonical, knowledge-based, and hand-labeled forms. Table 3 shows the three transcriptions of a Korean utterance: “그래서@여러가지로@의미가@깊은@달이기@때문입니다,” where @ to indicate a word boundary. First, this sentence is romanized as “g U l E s v @ jv l v g a z i l o @ xi m i g a @ g i p U n @ d a l i g i @ D E m u n i b n i d a.” The first three rows in the table are the reference phoneme sequences obtained canonical, knowledge-based, and hand-labeled transcriptions, respectively.

The canonical transcription shown in the first row of Table 3 is obtained from the Romanization. It can be seen that the only difference between the canonical form and the Romanization is ‘p’, as in ‘gipUn (깊은).’ This is because Jongseong can be pronounced as ‘ㄱ’, ‘ㄴ’, ‘ㄷ’, ‘ㄹ’, ‘ㅁ’, ‘ㅂ’, and ‘ㅇ’, they are romanized as ‘g’, ‘n’, ‘d’, ‘l’, ‘m’, ‘b’, and ‘N’, respectively. Hence, the Romanization ‘p’ is mapped to ‘b’. Moreover, the Korean consonants ‘ㄺ’, ‘ㄳ’, ‘ㄻ’, ‘ㄼ’, ‘ㅄ’, ‘ㄵ’, ‘ㄶ’, ‘ㅀ’, and ‘ㄾ’ [2], which should appear in both the Korean texts and Romanization transcriptions, are not included in the canonical form because they are not elements of the phoneme set described in Table 1.

[2] Romanization transcribes these Korean consonants as ‘lg’, ‘gs’, ‘lm’, ‘lb’, ‘bs’, ‘nz’, ‘nh’ and ‘lt’, respectively. These transcriptions are actually mapped such that {‘lg’,’gs’}, {‘lm’}, {‘lb’, ‘bs’}, {‘nz’, ‘nh’}, and {‘lh’, ‘lt’} in the Romanization are mapped into a single phone of ‘g’, ‘m’, ‘b’, ‘n’, and ‘l’ in the canonical form, respectively.

```
L1 in {n,jv}: g (22.0/7.0)
L1 in
{g,G,d,D,l,m,b,B,s,S,0,z,Z,c,k,t,p,h,E,ja,jE,v,e,
je,o,wa,wE,O,jo,u,wv,we,wi,ju,U,xi,i,N,gs,nz,n
h,lg,lm,lb,ls,lt,lp,lh,bs}: k (30.0/17.4)
L1 in {a,@}:
|   R1 in {a,o}: k (44.0/18.8)
|   R1 in {v,U}: g (180.0/76.1)
|   R1 in
{g,G,n,d,D,l,m,b,B,s,S,0,z,Z,c,k,t,p,h,E,ja,
jE,e,jv,je,wa,wE,O,jo,u,wv,we,wi,ju,xi,i,N,
gs,nz,nh,lg,lm,lb,ls,lt,lp,lh,bs,@}: k (6.0/4.3)
```

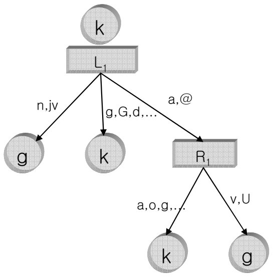


Figure 2: Example of decision tree building to derive pronunciation variation rules for a phone ‘k.’

On the other hand, the knowledge-based transcription is performed by using the phonological rules provided by the Korean pronunciation rulebook [15]. Here, the total number of phonological rules is 90. For example, the last word in the example, ‘때문입니다’, is romanized as ‘D E m u n i b n i d a,’ as shown in Table 3. However, it should be pronounced ‘때무님니다’, equivalently romanized as ‘D E m u n i m n i d a’, as expressed by the rule

$$?\text{-}?\text{-}b\text{+}n\text{+}? \rightarrow ?\text{-}?\text{-}m\text{+}n\text{+}? \qquad (2)$$

where the symbol ‘?’ stands for any phoneme shown in Table 1.

The hand-labeled transcription is done by a human expert. Thus, each utterance in the F-Korean01 and CleanSent01 databases could be transcribed using the three different transcription methods.

As a second step of the proposed method, we obtain a list of phoneme sequences after recognizing all the utterances in the development set of the CleanSent01 and F-Korean01 databases. From the list, we identify 42 patterns by comparing the recognized phoneme sequences and their corresponding transcriptions, where we use the canonical transcription for each utterance. Table 4 shows the rule patterns obtained from the sentence of Table 3 by using Eq. (1). It should be noted that a rule pattern is removed from the list of rule patterns if more than half of the neighboring phonemes are different. For example, one rule pattern in Table 4, d-a-l+i+g→l, is removed since ‘d,’ ‘a,’ and ‘i’ in the reference transcription are different. It is found that half of the wrong neighboring phonemes could be considered as errors incurred by the phoneme recognizer, and thus they are not used in the construction of the decision tree described in the next subsection

4.2. Deriving rules using a decision tree and adapting a dictionary

In this paper, a decision tree for each phoneme is constructed by using C4.5 [19] to derive the pronunciation variant rules. In order to adapt a dictionary for non-native speech recognition, a decision tree is first constructed from the rule patterns defined in Eq. (1). Since the C4.5 decision tree is characterized by multiple leaf nodes, the attributes used for training the C4.5 decision tree are the two left phonemes, L_1 and L_2, and the two right phonemes, R_1, and R_2. The output value of the decision tree is the target phoneme Y.

Fig. 2 shows an example of a decision tree constructed by using C4.5 and used to derive pronunciation variation rules for a center phoneme ‘k’. The fixed sets of attributes are L_1, L_2, R_1, and R_2, and each attribute corresponds to one of the 40 phonemes shown in Table 1. The output values can be several out of the 40 phonemes. For a given center phoneme ‘k’, the output value can change from ‘k’ to ‘g’ depending on L_1 and R_1. In other words, the output value can become ‘g’ instead of ‘k’ if L_1 is ‘n’ or ‘jv’. However, the output value changes to ‘g’ if R_1 is ‘v’ or ‘U’ when L_1 is ‘n’ or ‘jv’. The above procedure is applied to all the phonemes, resulting in a total of 40 decision trees.

Next, each decision tree is converted into an equivalent set of rules by tracing each path in the decision tree from the root node to each leaf node. For example, the decision tree shown in Fig. 2 can be converted into the following set of rules:

$$\begin{aligned} &\text{Rule } N\text{:} \\ &\quad R_1 = \text{‘v’} \\ &\qquad \rightarrow \text{class ‘g’ [Rule Accuracy]} \\ &\text{Default :} \\ &\qquad \text{class ‘k’} \end{aligned} \qquad (3)$$

where N is the rule number and N=1 in this example, and [Rule Accuracy] is the relative frequency of the rule applied to all the rule patterns associated with the center phoneme ‘k.’ If there is no rule for a rule pattern, the default rule is applied. After collecting all the rules obtained from the 40 decision trees, we apply a pruning technique to select the most effective rules. A rule is declared effective if the rule accuracy is greater than a given threshold. In this paper, the threshold is set to 25%.

Finally, the pruned rules are applied to adapting each reference dictionary described in Section 2.3, and the adapted dictionaries are then used for native and non-native ASR. For example, the Korean word ‘커지다’, which means ‘larger’, has a canonical transcription, ‘k v z i d a.’ If the rule in Eq. (3) is applied to the first phoneme ‘k’, the phoneme is changed into a variant phoneme ‘g’ because R_1 of ‘k’ is ‘v.’ Therefore, a pronunciation variant, ‘커지다: g v z i d a’, is added to the reference dictionary. As a result, the newly adapted dictionary includes the two elements ‘커지다: k v z i d a’ and ‘커지다: g v z i d a.’

5. SPEECH RECOGNITION EXPERIMENTS

In this section, we evaluated the performance of an ASR system using the dictionary derived by the proposed method and compared it to those of the systems using the reference dictionaries. The reference dictionaries used in this paper were canonical, knowledge-based and hand-labeled dictionaries, and the rules for dictionary adaptation were obtained by using three different database such as a non-native database only, a native database only, and the combination of native and non-native databases, denoted as ‘Non-native Rule’, ‘Native Rule’, and ‘Combined Rule’, respectively.

Table 5 shows the performance of the ASR system for non-native and native speech when the adapted dictionaries were applied. The first and second rows of the table show the performance when the dictionaries were adapted by using Non-native Rule and Native Rule, respectively. The WERs in this table were then compared with those shown in Table 2. The ASR system employing the reference dictionary constructed by canonical transcription gave WERs of 28.33% and 43.47% for non-native speech and native speech, respectively, as shown in the first row of Table 2. However,

Table 5: Comparison of the average word error rate (%) of the non-native and native ASR systems employing the dictionaries adapted by either non-native rules or native rules.

Dictionary adaptation by	Transcription for dictionary	Evaluation set	
		Non-native	Native
Non-native Rule	Canonical	22.87	46.65
	Knowledge-based	22.40	36.19
	Hand-labeled	22.33	34.94
Native Rule	Canonical	24.73	39.03
	Knowledge-based	24.80	34.66
	Hand-labeled	24.40	34.43

Table 6: Comparison of the average word error rate (%) of the non-native and native ASR systems employing the dictionaries adapted by the combination of non-native rules and native rules.

Dictionary adaptation by	Transcription for dictionary	Evaluation set	
		Non-native	Native
Combined Rule (Non-native + Native)	Canonical	22.40	39.49
	Knowledge-based	23.53	35.40
	Hand-labeled	22.60	34.60

the ASR system using the adaptive dictionary using Non-native Rule could achieve a relative WER reduction of 19.3% for non-native speech, but it had slightly worse performance for native speech compared with that of the reference dictionary. This was because the dictionary was adapted only by the rules from a non-native development database. On the other hand, this problem occurred in a reverse way such that the WER for native speech decreased but that for non-native speech marginally increased. We could find a similar tendency for the dictionaries obtained by knowledge-based and hand-labeled transcriptions. This motivated us to adapt the dictionaries by using the rules obtained from both non-native rules and native rules, which resulted in improved performance for both non-native and native speech, as shown in Table 6. Among the different dictionaries, we could achieve the best ASR performance when the dictionary transcribed by hand-labelers was adapted. Compared with the WER of the ASR system using the reference hand-labeled dictionary, the relative WER reductions were 18.5 % and 1.1% for non-native speech and native speech, respectively.

6. CONCLUSION

In this paper, we proposed a pronunciation variation modeling method for non-native and native speech recognition. The proposed method constructed rule patterns from native and non-native speech databases using an indirect data-driven approach, and applied the rules to adapt a dictionary to improve the performance of non-native and native speech recognition. It was shown from continuous non-native speech recognition experiments that the non-native ASR system using the dictionary adapted by the proposed method achieved the average WER reduction of 18.5%, compared to that using the baseline dictionary. Moreover, for native speech, the ASR system using the adapted dictionary also reduced the average WER by 1.1% compared to that using the baseline dictionary.

7. ACKNOWLEDGEMENT

This work was supported by the Korean Research Foundation Grant funded by the Korean Government (KRF-2007-314-D00245) and by the Ministry of Information and Communication through the Realistic Broadcasting Research Center (RBRC) at Gwangju Institute of Science and Technology.

8. REFERENCES

[1] S. Goronzy, M. Sahakyan, and W. Wokurek, “Is non-native pronunciation modeling necessary?” in *Proc. of Eurospeech*, vol. 1, pp. 309-312, Sept. 2001.

[2] J. Bellegarda, “An overview of statistical language model adaptation,” in *Proc. of ITRW on Adaptation Methods for Speech Recognition*, pp. 165-174, Aug. 2001.

[3] H. Strik and C. Cucchiarini, “Modeling pronunciation variation for ASR: A survey of the literature,” *Speech Comm*., vol. 29, nos. 2-4, pp. 225-246, Nov. 1999.

[4] S. Downey and R. Wiseman, “Dynamic and static improvements to lexical baseforms”, in *Proc. of ISCA Workshop on Modeling Pronunciation Variation*, pp. 157-162, May 1998.

[5] G. Tajchman, E. Fosler, and D. Jurafsdy, “Building multiple pronunciation models for novel words using exploratory computational phonology,” in *Proc. of Europseech*, pp. 2247-2250, Sept. 1995.

[6] E. Fosler-Lussier, “Multi-level decision trees for static and dynamic pronunciation models,” in *Proc. of Eurospeech*, pp. 463-466, Sept. 1999.

[7] M. Wolff, M. Eichner, and R. Hoffmann, “Automatic learning and optimization of pronunciation dictionaries,” in *Proc. of ITRW on Adaptation Methods for Speech Recognition*, pp. 159-162, Aug. 2001.

[8] I. Amdal, F. Korkmazasky, and A. C. Suredan, “Data-driven pronunciation modelling for non-native speakers using association strength between phones,” in *Proc of ASRU*, vol. 1, pp. 85-90, Aug. 2000.

[9] S. Goronzy, S. Rapp, and R. Kompe, “Generating non-native pronunciation variants for lexicon adaptation,” *Speech Comm*., vol. 42, no. 1, pp. 109-123, Sept. 2003.

[10] T. Svendsen, “Pronunciation modeling for speech technology,” *in Proc. of SPCOM*, pp. 11-16, Dec. 2004.

[11] B. W. Kim, D. L. Choi, Y. I. Kim, K. H. Lee, and Y. J. Lee, “Current state and future plans at SiTEC for speech corpora for common use,” *Malsori*, vol. 46, pp. 175-186, June 2003.

[12] B. W. Kim and Y. J. Lee, “Current states and future plans for speech corpora at SiTEC,” in *Proc. of the Acoustical Society of Korea Conference*, pp. 49-52, Sept. 2002.

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

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

[15] The Ministry of Education of Korea, *Korean Grammar Rulebook*, Daehan Printing & Publishing Co. Ltd., Seoul, Korea, 1995.

[16] J. J. Humphries, P. C. Woodland, and D. Pearce, “Using accent-specific pronunciation modeling for robust speech recognition,” in *Proc. of ICSLP*, pp. 2324-2327, Oct. 1996.

[17] J. Kim, C. Wang, M. Peabody, and S. Seneff, “An interactive English pronunciation dictionary for Korean learners,” in *Proc. of Interspeech*, pp. 1677-1680, Oct. 2004.

[18] T. Hain, “Implicit modelling of pronunciation variation in automatic speech recognition,” *Speech Comm.*, vol. 46, no. 2, pp. 171-188, June 2005.

[19] E. Fosler-Lussier, “A tutorial on pronunciation modeling for large vocabulary speech recognition,” *Lecture Notes in Artificial Intelligence*, vol. 2705, pp. 38-77, Apr. 2003.

[20] http://www2.cs.uregina.ca/~dbd/cs831/index.html.