PAPER Special Section on Processing Natural Speech Variability for Improved Verbal Human-Computer Interaction

# A Hybrid Acoustic and Pronunciation Model Adaptation Approach for Non-native Speech Recognition

**Yoo Rhee OH**[†], *Nonmember and* **Hong Kook KIM**[†a)], *Member*

**SUMMARY** In this paper, we propose a hybrid model adaptation approach in which pronunciation and acoustic models are adapted by incorporating the pronunciation and acoustic variabilities of non-native speech in order to improve the performance of non-native automatic speech recognition (ASR). Specifically, the proposed hybrid model adaptation can be performed at either the state-tying or triphone-modeling level, depending at which acoustic model adaptation is performed. In both methods, we first analyze the pronunciation variant rules of non-native speakers and then classify each rule as either a pronunciation variant or an acoustic variant. The state-tying level hybrid method then adapts pronunciation models and acoustic models by accommodating the pronunciation variants in the pronunciation dictionary and by clustering the states of triphone acoustic models using the acoustic variants, respectively. On the other hand, the triphone-modeling level hybrid method initially adapts pronunciation models in the same way as in the state-tying level hybrid method; however, for the acoustic model adaptation, the triphone acoustic models are then re-estimated based on the adapted pronunciation models and the states of the re-estimated triphone acoustic models are clustered using the acoustic variants. From the Korean-spoken English speech recognition experiments, it is shown that ASR systems employing the state-tying and triphone-modeling level adaptation methods can relatively reduce the average word error rates (WERs) by 17.1% and 22.1% for non-native speech, respectively, when compared to a baseline ASR system.

***key words:*** *non-native speech recognition, pronunciation variability, acoustic model adaptation, pronunciation model adaptation, state-tying level hybrid adaptation, triphone-modeling level hybrid adaptation*

## 1. Introduction

Due to the increasing use of non-native speech, such as in international business meetings or travel, both native and non-native speech commonly exist in a natural environment. Therefore, it becomes likely that non-native speech will be used in automatic speech recognition (ASR) systems; however, the ASR performance for non-native speech tends to degrade severely since typical ASR systems are only trained with native speech. Moreover, non-native speech displays broad variabilities due to a speaker's lack of fluency and/or the different pronunciation spaces between the target language and a speaker's mother tongue [1]. In attempts to handle non-native speech on ASR systems, research pertaining to these systems can be classified as acoustic modeling, pronunciation modeling, language modeling, and hybrid modeling approaches [2]–[11]. First, acoustic modeling approaches can simply retrain the acoustic models of native speech using a non-native speech database in order to compensate for the variability of non-native speech. However, there is a general lack of non-native speech data; instead, transforming or adapting acoustic models based on an analysis of pronunciation variation has been proposed [2]–[4]. Second, pronunciation modeling approaches build a multiple pronunciation dictionary by including pronunciation variants for each word in the pronunciation dictionary dedicated to non-native speech [2], [5]. To this end, several methods have been proposed that used either phoneme recognition or decision trees [6]–[9]. Third, language modeling approaches adjust a language model designed from native speech to compensate for the variability regarding the grammar or the speaking style of non-native speakers [10]. Finally, hybrid modeling approaches combine the above approaches to further improve the performance of non-native ASR [11].

In this paper, we focus on a hybrid modeling approach that combines pronunciation and acoustic model adaptations in order to handle the variability of non-native speech in an ASR system. In particular, we propose two types of hybrid modeling; one at the state-tying level and the other at the triphone-modeling level. The two hybrid modeling approaches first investigate pronunciation variabilities of non-native speech in an indirect data-driven manner and then classify each pronunciation variability as either pronunciation or acoustic variants. In other words, phoneme recognition is first performed to obtain $N$-best phoneme sequences using a development set of non-native speech. Then, the recognized $N$-best phoneme sequences are applied to a decision tree in order to derive the pronunciation variant rules [12]. Next, each pronunciation variant rule is classified into either a pronunciation variant or an acoustic variant, where they are used for pronunciation and acoustic adaptations, respectively.

After that, the state-tying level hybrid method builds a multiple pronunciation dictionary for non-native speech so that the pronunciation variants of each word are included in the pronunciation dictionary [12]; the states of the triphone acoustic models are clustered based on a decision tree using the acoustic variants [13]. In addition, the triphone-modeling level hybrid method adapts pronunciation models in the same way as the state-tying level hybrid method. However, there is one main difference between the state-tying level and the triphone-modeling level hybrid methods. In the triphone-modeling level method, the acoustic models are adapted by re-estimating the triphone acoustic models

Manuscript received November 30, 2009.
Manuscript revised February 26, 2010.

[†]The authors are with the School of Information and Communications, Gwangju Institute of Science and Technology (GIST), 1 Oryong-dong, Buk-gu, Gwangju 500–712, Korea.
a) E-mail: hongkook@gist.ac.kr

DOI: 10.1587/transinf.E93.D.2379

using the adapted pronunciation models and the states of the re-estimated triphone acoustic models are then clustered using the acoustic variants. In this paper, the proposed hybrid modeling methods are applied to a baseline ASR system constructed using a native-English speech database, where the English utterances spoken by Koreans are recognized as non-native speech.

The organization of the remainder of this paper is as follows. Following this introduction, the speech databases and a baseline ASR system are briefly described and then the effect of non-native speech variabilities on the ASR performance is discussed in Sect. 2. Section 3 describes how to investigate pronunciation variability including the derivation of variant rules from non-native speech and the classification of pronunciation or acoustic variants. In Sect. 4, we propose the state-tying and triphone-modeling level hybrid methods by combining the acoustic and pronunciation model adaptations. In Sect. 5, we evaluate the performance of a non-native ASR system employing the proposed hybrid model adaptation methods and compare their performance with that of the baseline ASR. Finally, we conclude our findings in Sect. 6.

## 2. Speech Database and Baseline ASR

In this section, we briefly describe the speech databases and a baseline ASR system that is trained with a native speech database. We then discuss the effect of non-native speech variability on the performance of this ASR system.

### 2.1 Speech Database

Two speech databases are used in this paper to train a baseline ASR system and to develop or evaluate a non-native speech ASR system. Table 1 provides the descriptions of the training set, in which a subset of the Wall Street Journal database (WSJ0) is used to train the system. WSJ0 is a 5,000-word closed-loop task used to evaluate the performance of a large vocabulary continuous ASR system [14]. The WSJ0 training set consists of 7,138 sentences (130,507 words in total) uttered by 83 native English speakers recorded using a Sennheiser close-talking microphone and several far-field microphones, where each utterance is sampled at a rate of 16 kHz.

In order to construct a non-native ASR system from the baseline ASR system, we use a subset of the Korean-spoken English corpus (K-SEC) [16] supported by the Speech Information Technology and Industry Promotion Center (SiTEC) in Korea. Table 2 describes the development set and the evaluation set, respectively, in several aspects. As shown in the table, the development set is composed of 1,103 isolated word utterances spoken by one Korean speaker, and the evaluation set is composed of 784 sentence utterances (8,176 words in total) with an average of 10.4 words per sentence, spoken by 49 Koreans and 7 native speakers. Moreover, the isolated word and sentence utterances are the read speech of phonetically balanced words (PBW) and a part of one of Aesops Fables, *The Wind and the Sun*, respectively.

**Table 1** Comparison of the training corpus.

| Item | Training set |
|---|---|
| Language | American English |
| Database | WSJ0 [14] |
| Mother tongue | American English |
| Utterance type | Sentence |
| No. of speakers | 83 |
| No. of sentences | 7,138 |
| No. of words | 130,507 |
| Amount of speech data | 3 hours |

### 2.2 Baseline ASR System for Native Speech

As the speech feature vector of an ASR system, we extract 12 mel-frequency cepstral coefficients (MFCCs) with logarithmic energy for every 10 ms analysis frame and concatenate their first and second derivatives, resulting in a 39-dimensional feature vector. For all training and test utterances, we apply cepstrum mean normalization and energy normalization to the feature vectors. For the acoustic models, cross-word triphone hidden Markov models (HMMs) are constructed based on 3-state left-to-right, context-dependent HMMs with 4-mixture Gaussian distributions, and they are trained using the HTK version 3.2 Toolkit [17].

Figure 1 shows the main procedure for constructing acoustic models and building pronunciation models for the baseline ASR system. In other words, all of the triphone models are expanded from 41 monophones, which also include silence and pause models, as shown in Table 3. After that, the states of the triphone models are joined using a decision tree [18]. As a result, the acoustic models consist of 9,655 triphones and 5,297 states, referred to as AM0 throughout this paper. Moreover, each pronunciation of a word is built from the Carnegie Mellon University (CMU) pronunciation dictionary [19] and any words missing in the CMU dictionary are manually transcribed. A set of these baseline pronunciation models with 87 unique words and 340 pronunciations is referred to as PM0.

In addition, in order to explore discrepancies in the behavior of the pronunciation and the acoustic models due to

**Table 2** Comparison of the development and evaluation set for non-native ASR.

| Item | Development set | Evaluation set | |
|---|---|---|---|
| Language | American English | | |
| Database | K-SEC [16] | | |
| Mother tongue | Korean | American English | Korean |
| Utterance type | Isolated word | Sentence | |
| No. of speakers | 1 | 7 | 49 |
| No. of sentences | - | 98 | 686 |
| No. of words | 1,103 | 1,022 | 7,154 |
| Amount of speech data | 0.3 hours | 0.2 hours | 1.1 hours |

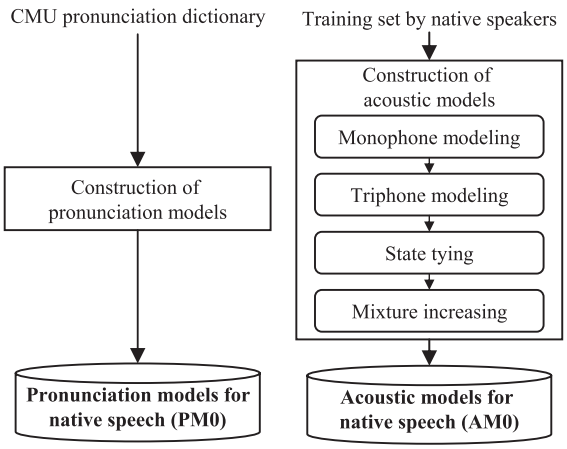


**Fig. 1** Main procedure for constructing acoustic models and pronunciation models in a baseline ASR system.

**Table 3** List of phonemes used for a baseline ASR system.

| | |
|---|---|
| Vowels (15) | AA AE AH AO AW AY EH ER EY IH IY OW OY UH UW |
| Consonants (24) | B CH D DH F G HH JH K L M N NG P R S SH T TH V W Y Z ZH |
| Silences (2) | sil sp |

**Table 4** Comparison of average WERs (%) of a baseline ASR system for non-native and native speech.

| | Native | Non-native |
|---|---|---|
| Baseline ASR system | 0.7 | 19.9 |

differences between the target language and the speaker's mother tongue, the texts collected from the evaluation set are only used to construct a backed-off bigram language model with a perplexity of 3. It is noted here that there is no out-of-vocabulary word. However, if the language model is generated using a large text corpus and thus the language model perplexity is increased, the performance improvement may be different from that obtained in this paper [15].

### 2.3 Effect of Non-native Speech Variability on the ASR Performance

Table 4 compares the average word error rates (WERs) of the baseline ASR system for the two evaluation sets described in Table 2. As shown in the table, the baseline ASR system achieves average WER of 0.7% for native speech. Note here that the high recognition performance is due to the small size of the language model. On the other hand, average WER for non-native speech is severely increased at 19.9%. This is because the baseline ASR system is trained with a native speech database thus it cannot properly handle non-native speech that yields a broad range of speech variability against native speech. As mentioned in Sect. 1, non-native speech has variabilities that are originated from either a speaker's lack of fluency or from the difference in pronunciation spaces between the mother tongue and the target language. In addition, such variabilities also occur due to coarticulation effects within a specific context. These issues are discussed in detail in the next section.

## 3. Pronunciation Variability of Non-native Speech

In this section, we first describe how to obtain the pronunciation variant rules from a non-native speech database in a data-driven approach. Next, we describe how to classify each rule into either a pronunciation or an acoustic variant rule, which is used for pronunciation adaptation or acoustic adaptation, respectively.

### 3.1 Analysis of Pronunciation Variant Rules from Non-native Speech

The pronunciation variant rules of non-native speech are obtained in an indirect data-driven approach based on a decision tree [12], which is performed in three steps. For a given utterance, $N$-best phoneme sequences are obtained using phoneme recognition and aligned using dynamic programming with the reference sequence of the utterance. After that, pronunciation variant rules are derived based on decision trees.

#### 3.1.1 Phoneme Recognition

We first perform phoneme recognition using each utterance in the development set of non-native speech. As a phoneme recognizer, we use acoustic models of the baseline ASR system and a back-off bigram language model that is trained with the reference phoneme sequences of the training text data. By performing phoneme recognition, we obtain $N$-best phoneme sequences for each utterance instead of a 1-best phoneme sequence in order to get the meaningful pronunciation variabilities. In this paper, we actually use the 200-best phoneme sequences; the phoneme recognition accuracies for the 1-best and 200-best phoneme sequences are measured as 18.5% and 61.3%, respectively.

#### 3.1.2 Alignment Using Dynamic Programming

For each utterance, the recognized $N$-best phoneme sequences are aligned using a dynamic programming (DP) algorithm, where the reference phoneme sequence is obtained from the CMU pronunciation dictionary [19]. Table 5 shows an example of the alignment result for an utterance, 'unchained melody,' where the reference phoneme sequence and one of the recognized $N$-best phoneme sequence are /AH N CH EY N D M EH L AH D IY/ and /T AH N S AH N D AH L AH T IY/, respectively[†]. In the table, /@/ in the upper row of the alignment result indicates an insertion error but /@/ in the lower row indicates a deletion error.

Among the $N$ alignment results for each utterance, we then select the $M$ ($M < N$) best-matched results as an alter-

[†]From now on, all pronunciation symbols are denoted in the form of the two-letter uppercase ARPAbet [20], as it is usual to use ARPAbet symbols in speech recognition.

**Table 5** Example of a DP-based alignment between the recognized phoneme sequence and the reference phoneme sequence for an utterance, 'unchained melody.' The symbol /@/ in the alignment result indicates an insertion or a deletion error.

| Text | Unchained melody | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Reference phoneme sequence | AH N CH EY N D M EH L AH D IY | | | | | | | | | | | | |
| Recognized phoneme sequence | T AH N S AH N D AH L AH T IY | | | | | | | | | | | | |
| Alignment result | @ | AH | N | CH | EY | N | D | M | EH | L | AH | D | IY |
| | T | AH | N | S | AH | N | D | @ | AH | L | AH | T | IY |

**Table 6** Example of the pronunciation rule patterns obtained from the alignment result in Table 4.

| | |
|---|---|
| @-@-@+AH+N→T | N-D-M+EH+L→@ |
| @-@-AH+N+CH→AH | D-M-EH+L+AH→AH |
| @-AN-N+CH+EY→N | M-EH-L+AH+D→L |
| AN-N-CH+EY+N→S | EH-L-AH+D+IY→AH |
| N-CH-EY+N+D→AH | L-AH-D+IY+@→T |
| CH-EY-N+D+M→N | AH-D-IY+@+@→IY |
| EY-N-D+M+EH→D | |

native to the compensation of the phoneme mis-recognition. Next, the pronunciation rule patterns for each selected alignment result are obtained in the form of

$$L_2 - L_1 - X + R_1 + R_2 \ \rightarrow \ Y \tag{1}$$

where /$X$/ is the target 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. Table 6 shows the pronunciation rule patterns obtained from the alignment result in Table 5. In other words, each pair in the alignment result has an independent pronunciation rule pattern defined in the form of Eq. (1). For example, the pair (/@/, /T/) has the pronunciation rule pattern /@-@-@+AH+N/→/T/; other pronunciation rule patterns can then be generated in the same manner.

### 3.1.3 Derivation of Pronunciation Variant Rules

In order to derive the pronunciation variant rules from the development set of non-native speech, a decision tree for each target phoneme is generated using all the corresponding pronunciation rule patterns. In this paper, C4.5 is used to generate a decision tree, which is a software extension of the basic ID3 algorithm designed by Quinlan [21]. In the decision tree, the attributes of the decision tree for a target phoneme /$X$/ are the two left phonemes, /$L_1$/ and /$L_2$/, and the two right phonemes, /$R_1$/ and /$R_2$/. In addition, the output class is the phoneme /$Y$/ mapped from the target phoneme /$X$/. Next, each decision tree is converted into an equivalent set of pronunciation variant rules using C4.5, which is represented as

**Table 7** Example of a decision tree and the derived pronunciation rule set for the target phoneme /Z/.

| A decision tree for /Z/ | Derived pronunciation variant rule set for /Z/ |
|---|---|
| $R_1$ = @: Z | Rule 1: |
| $R_1$ = AH: Z | $L_1$ = IH, $R_1$ = IY |
| $R_1$ = AY: Z | → class S |
| $R_1$ = EH: Z | Rule 2: |
| $R_1$ = ER: Z | $R_1$ = @ |
| $R_1$ = IH: Z | → class Z |
| $R_1$ = OW: Z | Rule 3: |
| $R_1$ = IY: | $R_1$ = AH |
| \| $L_1$ = @: S | → class Z |
| \| $L_1$ = AH: S | |
| \| $L_1$ = AW: S | Default class : Z |
| \| $L_1$ = AY: S | |
| \| $L_1$ = ER: S | |
| \| $L_1$ = EY: S | |
| \| $L_1$ = IH: S | |
| \| $L_1$ = IY: S | |
| \| $L_1$ = OW: S | |
| \| $L_1$ = OY: Z | |
| \| $L_1$ = R: S | |
| \| $L_1$ = UW: S | |

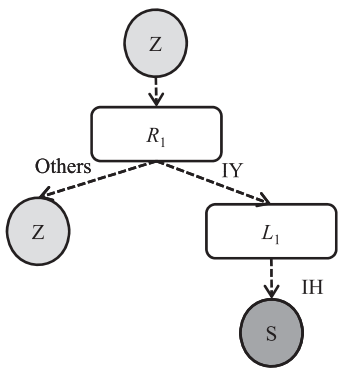


**Fig. 2** Illustration of the simplified decision tree for the target phoneme /Z/ according to the pronunciation variant rules listed in Table 6.

$$\begin{aligned} &\text{Rule } id: \\ &\quad L_2 = p_1, L_1 = p_2, R_1 = p_3, R_2 = p_4 \\ &\qquad \rightarrow \text{class } phoneme_{variant} \\ &\text{Default class : } phoneme_{default} \end{aligned} \tag{2}$$

where *id* is the identifier of a pronunciation variant rule, and '$L_2 = p_1, L_1 = p_2, R_1 = p_3, R_2 = p_4$' is a context in which *id* is applied. In other words, /$phoneme_{target}$/ is mapped into /$phoneme_{variant}$/ if the context has the form of /$p_1$-$p_2$-$phoneme_{target}$+$p_3$+$p_4$/; otherwise /$phoneme_{target}$/ is mapped into /$phoneme_{default}$/.

Table 7 shows an example of a decision tree and the derived pronunication variant rule set for the target phoneme /Z/, which is obtained from all the utterances in the development set. The derived rule set in Table 7 can be depicted as a simplified decision tree, which is shown in Fig. 2. In Table 7, ':' means a separator of a condition. In addition, 'x=y' and 'z' that are separated terms by ':' interpret as 'when x is y' and 'a mapped phoneme is z,' respectively. Moreover, '|' is

used when more than one conditions are applied.

### 3.2 Classification of Pronunciation Variant Rules for Non-native Speech

We classify each pronunciation variant rule into either a pronunciation variant or an acoustic variant based on the assumption that a pronunciation variant occurs due to the coarticulation effect within a specific context, while an acoustic variant occurs due to the difference in pronunciation spaces between the mother tongue and target language. For example, 'misstep' could be uttered by a Korean as /M IH S UW S T EH P/ instead of /M IH S S T EH P/ since Koreans tend to pronounce each word clearly by inserting a vowel even if the /S/ in the sub-word 'mis' and the /S/ in the sub-word 'step' are adjacent in the context. Thus, the pronunciation variant rule, /S/→/S UW/, is classified as a pronunciation variant due to the coarticulation effect. As another example, 'strike' could be uttered by a Korean as /S UW T UW R AY K/ instead of /S T R AY K/ due to the different syllable structure of languages [22]. The situation is caused by the fact that 'strike' has a consonant cluster 'str' in the word; however, the syllable structure of Korean allows no more than one consonant and thus Koreans try to utter the complex consonants /S T R/ as /S UW T UW R/ by inserting a vowel /UW/ between each two concatenated consonants. Therefore, a pronunciation variant rule, /S T R/→/S UW T UW R/, is classified as a pronunciation variant due to a specific context for non-native speech. On the other hand, Koreans often utter 'five' as /P AY B/ instead of /F AY V/ because the Korean language does not include the two phonemes /F/ and /V/. Thus, the pronunciation variant rules, /F/→/P/ and /V/→/B/, are classified as acoustic variants for Korean speakers. In practice, we classify a pronunciation variant rule as an acoustic variant if the default class has a different phoneme as a target phoneme; otherwise, the pronunciation variant rule is classified as a pronunciation variant.

Table 8 shows an example of the pronunciation variant rule sets for the target phonemes /Z/ and /TH/, which are obtained from all the utterances in the development set. As shown in the table, /Z/ has a same default class but /TH/ has a different default class. Among the pronunciation variant rules, the pronunciation variant /TH/→/DH/ is classified as an acoustic variant and others are as pronunciation variants. In other words, we have 8 different pronunciation variants such as /*-IH-Z+IY+*/→/S/, /*-*-Z+@+*/→/Z/, /*-*-Z+AH+*/→/Z/, /*-*-TH+@+*/→/S/, /*-*-TH+AH+*/→/S/, /*-AH-TH+*+NG/→/IY/, /*-*-TH+AO+*/→/TH/, and /*-*-TH+IH+*/→/F/, where '*' indicates any phoneme.

After classifying the pronunciation variant rules obtained from all the utterances in the development set into pronunciation variants or acoustic variants, we finally obtain 473 pronunciation variants and four acoustic variants. Table 9 classifies the 473 pronunciation variants according to the form of pronunciation variant rules and shows the number of pronunciation variants classified to each form. Moreover, the four acoustic variants are /G/→/sil/, /L/→/R/, /TH/→/DH/, and /ZH/→/Z/. Note here that the phonemes /R/, /TH/, /DH/, and /ZH/ do not exist in the Korean phoneme set, thus the pronunciations are likely to be affected by the speaker's mother tongue.

**Table 8** Construction of the pronunciation variant rule sets for the target phonemes /Z/ and /TH/.

| Derived pronunciation variant rule set for /Z/ | Derived pronunciation variant rule set for /TH/ |
|---|---|
| Rule 1:<br>$L_1$ = IH, $R_1$ = IY<br>→ class S<br>Rule 2:<br>$R_1$ = @<br>→ class Z<br>Rule 3:<br>$R_1$ = AH<br>→ class Z<br><br>**Default class : Z** | Rule 1:<br>$R_1$ = @<br>→ class S<br>Rule 2:<br>$R_1$ = AH<br>→ class S<br>Rule 3:<br>$L_1$ = AH, $R_2$ = NG<br>→ class IY<br>Rule 4:<br>$R_1$ = AO<br>→ class TH<br>Rule 5:<br>$R_1$ = IH<br>→ class F<br>**Default class : DH** |

**Table 9** Classification of pronunciation variants according to their forms, where the number of a form means the number of pronunciation variants belonging to the form.

| Form | Number |
|---|---|
| $X+R_1$ | 113 |
| $X+*+R_2$ | 27 |
| $L_1-X$ | 88 |
| $L_2-*-X$ | 77 |
| $X+R_1+R_2$ | 23 |
| $L_1-X+R_1$ | 41 |
| $L_1-X+*+R_2$ | 22 |
| $L_2-L_1-X$ | 36 |
| $L_2-*-X+R_1$ | 27 |
| $L_2-*-X+*+R_2$ | 8 |
| $L_2-L_1-X+R_1$ | 8 |
| $L_1-X+R_1+R_2$ | 1 |
| $L_2-L_1-X+*+R_2$ | 1 |
| $L_2-*-X+R_1+R_2$ | 1 |

## 4. Hybrid Model Adaptation for Non-native Speech

In this section, we first introduce pronunciation and acoustic model adaptation methods. Then, we propose two hybrid pronunciation and acoustic model adaptations such as state-tying and triphone-modeling level hybrid methods for a non-native ASR system by incorporating the pronunciation and acoustic variants derived from the development set of native speech.

### 4.1 Pronunciation Model Adaptation

The pronunciation models employed in ASR systems are commonly generated based on the native pronunciations for

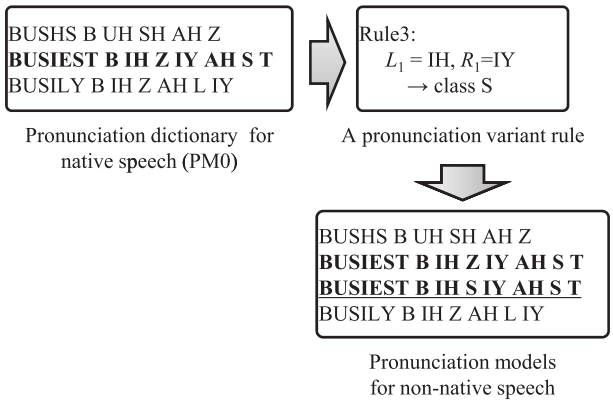


**Fig. 3** Example of building a multiple pronunciation dictionary for non-native speech by using the rules listed in Table 7.

each word. Therefore, we need to adapt pronunciation models in order to compensate for variabilities that are originated from non-native speech. In other words, a pronunciation sequence for each word is examined to determine whether or not it matches the pronunciation rules obtained in Sect. 3.2. If a matched pronunciation rule is found, a variant pronunciation sequence is generated by using this matched pronunciation rule and is added as an additional pronunciation for the corresponding word. This pronunciation model adaptation method is used as a first step for the proposed hybrid model adaptation methods.

Figure 3 shows an example of building a multiple pronunciation dictionary by using the pronunciation rules described in Table 8. For example, the pronunciation dictionary in the baseline ASR system has a pronunciation sequence of the word 'busiest' as /B IH Z IY AH S T/. However, it is found from the pronunciation variability analysis described in Sect. 3.2 that a pronunciation variant rule regarding /Z/ (Rule 1 in Table 8) is /*-IH-Z+IY+*/→/S/. Thus, a variant pronunciation sequence of the word is generated as /B IH S IY AH S T/ and is included in the pronunciation dictionary for non-native speech.

## 4.2 Acoustic Model Adaptation

To adapt the acoustic models for non-native speech, different procedures are applied to cluster the states of the triphone acoustic models depending on whether or not the central phone of a triphone has an acoustic variant [23]. For instance, for a phoneme with no acoustic variant, a decision tree for /$phoneme_{target}$/ is generated by using all the triphone acoustic models having the form of /*-$phoneme_{target}$+*/. On the other hand, for a phoneme having an acoustic variant such as /$phoneme_{target}$/→/$phoneme_{variant}$/, a decision tree for /$phoneme_{target}$/ is generated by using the triphone acoustic models having the form of either /*-$phoneme_{target}$+*/ or /*-$phoneme_{variant}$+*/. After clustering all acoustic models using the decision tree, the models in each leaf node of the decision tree are tied with their representative phonemes.

Figures 4 and 5 illustrate the procedure for generating a decision tree for /Z/ and /TH/, respectively, where /Z/ has no acoustic variants but /TH/ has the acoustic variant /TH/→/DH/ as described in Table 8.

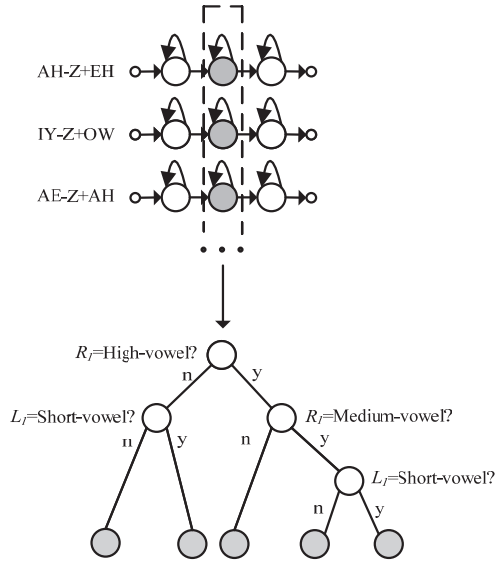


**Fig. 4** Example of generating a decision tree for the phoneme /Z/ without any acoustic variant.

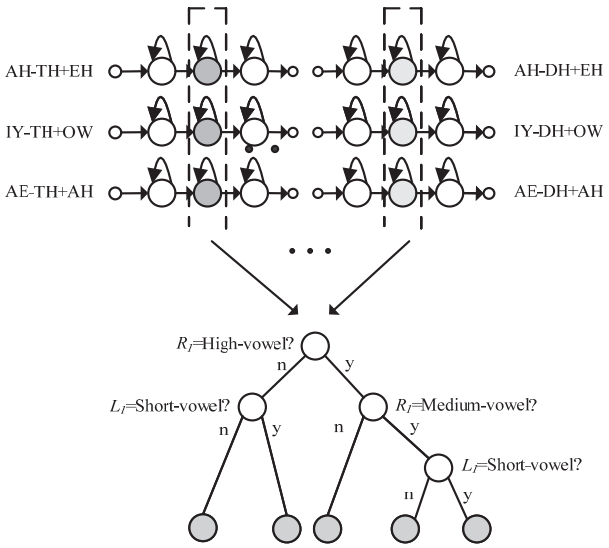


**Fig. 5** Example of generating a decision tree for the phoneme /TH/ having an acoustic variant of /TH/→/DH/.

## 4.3 Hybrid Model Adaptation

In this subsection, we propose two different hybrid model adaptation methods performed at the state-tying level or at the triphone-modeling level of acoustic model adaptation by using the pronunciation models obtained in Sect. 4.1. In other words, the state-tying level hybrid method adapts pronunciation models by including the pronunciation variants of each word in the pronunciation dictionary [12]. In addition, states of the triphone acoustic models are tied using

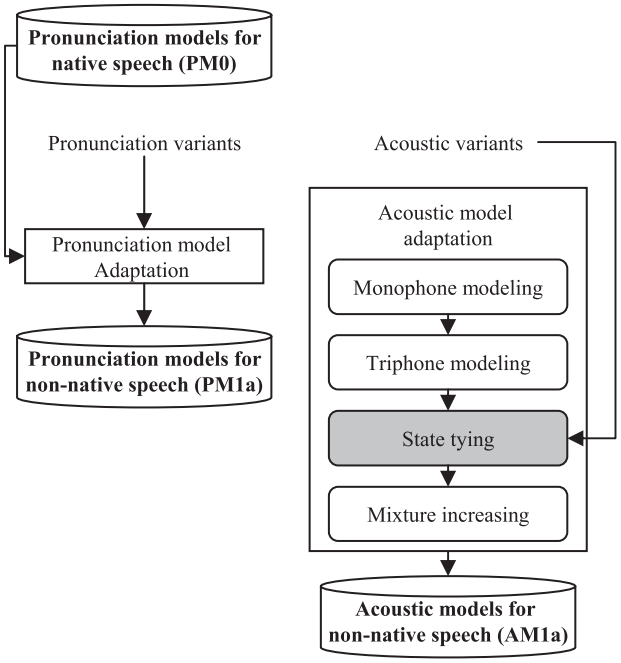


**Fig. 6** Main procedure of the proposed state-tying level hybrid model adaptation for non-native speech.

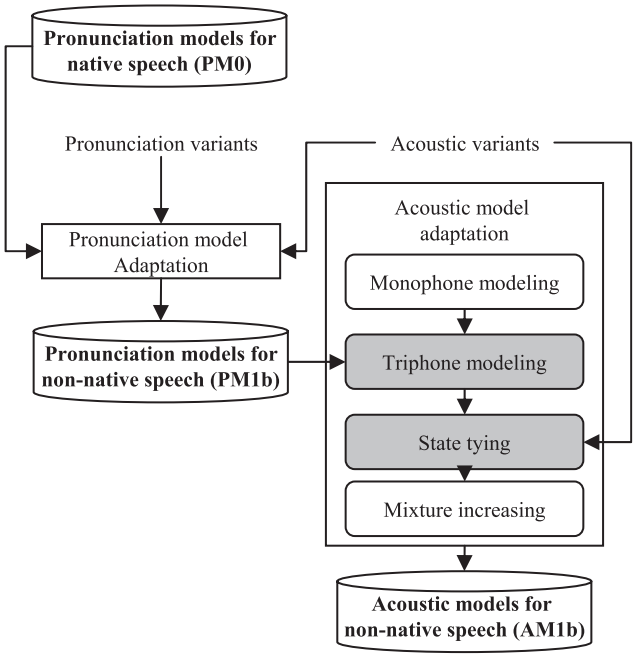


**Fig. 7** Main procedure of the proposed triphone-modeling level hybrid model adaptation for non-native speech.

the acoustic variants [13]. On the other hand, the triphone-modeling level hybrid method adapts pronunciation models in the same way as the state-tying level hybrid method, then re-estimates the triphone acoustic models using the adapted pronunciation models, and clusters the states of the re-estimated triphone acoustic models using the acoustic variants.

#### 4.3.1 State-Tying Level Hybrid Model Adaptation

Figure 6 shows the main procedure of the state-tying level hybrid model adaptation method. In the figure, pronunciation and acoustic model adaptations are performed according to the classification of pronunciation variant rules. As described in Sect. 4.1, the pronunciation model adaptation builds a multiple pronunciation dictionary for non-native speech using the pronunciation variants. Moreover, the acoustic model adaptation constructs acoustic models for non-native speech by re-tying the states of triphone acoustic models according to acoustic variants using the training set for native speech as described in Table 1.

#### 4.3.2 Triphone-Modeling Level Hybrid Model Adaptation

Figure 7 illustrates the procedure of the triphone-modeling level hybrid model adaptation method. As opposed to the state-tying level hybrid method, the pronunciation and acoustic model adaptations are combined. First, pronunciation model adaptation is performed using all the pronunciation variant rules in the same way as the state-tying level hybrid method. Next, acoustic model adaptation is performed using the pronunciation models for non-native speech and the acoustic variants. That is, the reference transcriptions of the training utterances for acoustic models are newly generated by forced-alignment using the pronunciation model adapted from non-native speech. Then, the triphone acoustic models of the baseline ASR system are re-estimated using the newly generated reference transcriptions. Finally, the states of the re-estimated triphone acoustic models are re-clustered in the same way as in the state-tying level hybrid method.

## 5. Speech Recognition Experiments

### 5.1 Acoustic and Pronunciation Model Adaptations

We first performed phoneme recognition using the baseline ASR system for the development set of non-native speech, and then obtained the 200-best phoneme sequences for each utterance. For each phoneme sequence, a DP based alignment was performed with the corresponding reference phoneme sequence in order to extract the 20 best-matched alignment results for each utterance. As previously mentioned, the selection of the 20 best-matched sequences out of the 200-best phoneme sequences was carried out to improve the accuracy in deriving the pronunciation variant rules. Next, the pronunciation variant rules were derived and subsequently classified into either pronunciation or acoustic variants. As a result, /G/→/sil/, /L/→/R/, /TH/→/DH/, and /ZH/→/Z/ were classified into acoustic variants and the others were considered pronunciation variants, as described in Sect. 3.2. After that, we obtained two sets of pronunciation models 1) using the pronunciation variants alone, which consisted of 448 pronunciations and was referred to

**Table 10** Comparison of average WERs (%) of the baseline ASR system and ASR systems employing acoustic and pronunciation model adaptations.

| ASR systems | Native | Non-native | Relative WER Reduction (for non-native only) |
|---|---|---|---|
| Baseline (AM0+PM0) | 0.7 | 19.9 | - |
| Pronunciation model adaptation using pronunciation variants (AM0+PM1a) | 0.7 | 18.2 | 8.5 |
| Pronunciation model adaptation using both acoustic and pronunciation variants (AM0+PM1b) | 0.7 | 17.3 | 13.1 |
| Acoustic model adaptation (AM1a+PM0) | 0.9 | 18.1 | 9.0 |

as PM1a, and 2) using both the pronunciation and acoustic variants, which consisted of 493 pronunciations and was referred to as PM1b. Note that AM0 and PM0 were the acoustic models and pronunciation models of the baseline ASR system, as mentioned in Sect. 2.2. Moreover, the acoustic model adaptation was applied by re-clustering the states of triphone acoustic models using the acoustic variants in the same way as in the state-tying level hybrid method; a set of the adapted acoustic models consisted of 5,361 states and 9,655 triphones and was referred to as AM1a.

Table 10 compares average WERs of the baseline ASR system and several ASR systems employing the adapted pronunciation models using the pronunciation variants (PM1a), the adapted pronunciation models using both the pronunciation and acoustic variants (PM1b), and the adapted acoustic models using the acoustic variants (AM1a). For PM1a and PM1b, we used the acoustic models of the baseline ASR system (AM0), whereas PM0 was used to evaluate the performance of AM1a. It was shown from the table that for non-native speech average WERs of ASR systems using the adapted pronunciation models were relatively reduced by 8.5% and 13.1%, respectively, even though the confusions of PM1a and PM1b are increased, when compared to the baseline ASR system. In addition, an ASR system only using the adapted acoustic models also gave similar WER to an ASR system only using the adapted pronunciation models, PM1a, but it degraded native ASR performance a little.

### 5.2 Performance Comparison of Hybrid Model Adaptation Methods

In this subsection, we compared the ASR performance of the state-tying and triphone-modeling level hybrid model adaptation methods. For the state-tying level hybrid method, acoustic models (AM1a) and pronunciation models (PM1a) were adapted using the acoustic and pronunciation variants, respectively, according to the classification of the pronunciation variant rules. However, for the triphone-modeling level hybrid method, pronunciation models were first adapted using both acoustic and pronunciation variants. After that, acoustic models were adapted using the adapted pronunciation models and the states of the triphone acoustic models were then re-clustered using the acoustic variants. A set of the acoustic models adapted by the triphone-modeling level hybrid method consisted of 6,376 states and 14,655 triphones and was referred to as AM1b.

**Table 11** Comparison of average WERs (%) of the baseline ASR system and ASR systems employing state-tying and triphone-modeling level hybrid model adaptations.

| ASR systems | Native | Non-native | Relative WER Reduction (for non-native only) |
|---|---|---|---|
| Baseline (AM0+PM0) | 0.7 | 19.9 | - |
| State-tying level hybrid model adaptation (AM1a+PM1a) | 0.8 | 16.5 | 17.1 |
| Triphone-modeling level hybrid model adaptation (AM1b+PM1b) | 0.9 | 15.5 | 22.1 |

Table 11 compares average non-native WERs of the baseline ASR system and those of two ASR systems employing the state-tying and triphone-modeling level hybrid model adaptation methods, respectively. It was shown from the table that ASR systems employing the state-tying level and the triphone-modeling level hybrid methods achieved average WERs of 16.5% and 15.5% for non-native speech, respectively, which corresponded to relative WER reductions of 17.1% and 22.1%, respectively, when compared to the baseline ASR system.

## 6. Conclusion

In this paper, we proposed two hybrid model adaptation methods by combining pronunciation and acoustic model adaptations in order to incorporate speech variability for non-native ASR. Specifically, acoustic model adaptation was applied at different levels of the acoustic modeling procedure, such as at the state-tying or triphone-modeling level. Both the state-tying and the triphone-modeling level hybrid methods first investigated pronunciation variability of non-native speech and then classified them as either pronunciation or acoustic variants. After that, the state-tying level hybrid method adapted the pronunciation and acoustic models by including the variant pronunciation sequence using pronunciation variants and by clustering states of the triphone acoustic models using the acoustic variants, respectively. On the other hand, the triphone-modeling level hybrid method first adapted pronunciation models in the same way as the state-tying level hybrid method; however, the method then re-estimated the triphone acoustic models using the adapted pronunciation models and clustered states of the re-estimated triphone acoustic models using the

acoustic variants. From the Korean-spoken English speech-recognition experiments, it was shown that the triphone-modeling level hybrid adaptation method was better than the state-tying level hybrid adaptation method, whereas the former incurred additional complexity in training. In addition, it was found that ASR systems employing the state-tying and the triphone-modeling level hybrid adaptation methods could relatively reduce average WERs for non-native speech by 17.1% and 22.1%, respectively, when compared to the baseline ASR system.

## Acknowledgments

This work was supported in part by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2007-314-D00245) and by the MKE (The Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2010-C1090-1021-0007).

### References

[1] D.V. Compernolle, "Recognizing speech of goats, wolves, sheep and ... non-natives," Speech Commun., vol.35, no.1, pp.71–79, Aug. 2001.
[2] R. Gruhn, K. Markov, and S. Nakamura, "A statistical lexicon for non-native speech recognition," Proc. ICSLP, pp.1497–1500, Jeju Island, Korea, Oct. 2004.
[3] S. Steidl, G. Stemmer, C. Hacker, and E. Noth, "Adaptation in the pronunciation space for non-native speech recognition," Proc. ICSLP, pp.2901–2904, Jeju Island, Korea, Oct. 2004.
[4] J. Morgan, "Making a speech recognizer tolerate non-native speech through Gaussian mixture merging," Proc. InSTIL/ICALL Symposium on Computer-Assisted Language Learning, paper 052, Venice, Italy, June 2004.
[5] A. Raux, "Automated lexical adaptation and speaker clustering based on pronunciation habits for non-native speech recognition," Proc. ICSLP, pp.613–616, Jeju Island, Korea, Oct. 2004.
[6] H. Strik and C. Cucchiarini, "Modeling pronunciation variation for ASR: A survey of the literature," Speech Commun., vol.29, nos.2-4, pp.225–246, Nov. 1999.
[7] E. Fosler-Lussier, "Multi-level decision trees for static and dynamic pronunciation models," Proc. Eurospeech, pp.463–466, Budapest, Hungary, Sept. 1999.
[8] I. Amdal, F. Korkmazaskiy, and A.C. Suredran, "Data-driven pronunciation modelling for non-native speakers using association strength between phones," Proc. ASR, vol.1, pp.85–90, Paris, France, Sept. 2000.
[9] S. Goronzy, S. Rapp, and R. Kompe, "Generating non-native pronunciation variants for lexicon adaptation," Speech Commun., vol.42, no.1, pp.109–123, Sept. 2003.
[10] J. Bellegarda, "An overview of statistical language model adaptation," Proc. ISCA Workshop on Adaptation Methods for Speech Recognition, pp.165–174, Sophia-Antipolis, France, Aug. 2001.
[11] G. Bouselmi and I. Illina, "Combined acoustic and pronunciation modelling for non-native speech recognition," Proc. Interspeech, pp.1449–1452, Antwerp, Belgium, Aug. 2007.
[12] M. Kim, Y.R. Oh, and H.K. Kim, "Non-native pronunciation variation modeling using an indirect data driven method," Proc. ASRU, pp.231–236, Kyoto, Japan, Dec. 2007.
[13] Y.R. Oh, J.S. Yoon, and H.K. Kim, "Acoustic model adaptation based on pronunciation variability analysis for non-native speech recognition," Speech Commun., vol.49, no.1, pp.59–70, Jan. 2007.
[14] D. Paul and J. Baker, "The design for the Wall Street Journal-based CSR corpus," Proc. DARPA Speech and Language Workshop, pp.357–362, Arden House, NY, Feb. 1992.
[15] D. Klakow and J. Peters, "Testing the correlation of word error rate and perplexity," Speech Commun., vol.38, nos.1–2, pp.19–28, Sept. 2002.
[16] S.-C. Rhee, S.-H. Lee, S.-K. Kang, and Y.-J. Lee, "Design and construction of Korean-spoken English corpus (K-SEC)," Proc. ICSLP, pp.2769–2772, Jeju Island, Korea, Oct. 2004.
[17] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, The HTK Book (for HTK Version 3.2), Microsoft Corporation, Cambridge University Engineering Department, Dec. 2002.
[18] S. Young, J. Odell, and P. Woodland, "Tree-based state tying for high accuracy acoustic modeling," Proc. ARPA Human Language Technology Workshop, pp.307–312, Princeton, NJ, March 1994.
[19] H. Weide, The CMU Pronunciation Dictionary, release 0.6, Carnegie Mellon University, 1998.
[20] J.R. Deller, J.H.L. Hansen, and J.G. Proakis, Discrete-Time Processing of Speech Signals, IEEE Press, New York, NY, 2000.
[21] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Mateo, CA, 1993.
[22] P. Ladefoged, A Course in Phonetics (Fourth Edition), Harcourt College Publishers, New York, NY, 2001.
[23] Y.R. Oh, M. Kim, and H.K. Kim, "Acoustic and pronunciation model adaptation for context-independent and context-dependent pronunciation variability of non-native speech," Proc. ICASSP, pp.4281–4284, Las Vegas, NV, April 2008.

**Yoo Rhee Oh** received a B.S. degree in Computer Science from Kyungpook National University, Korea in 2004, and an M.S. degree in Information and Communications Engineering from the Gwangju Institute of Science and Technology (GIST), Korea in 2006. She is now a Ph.D. student at GIST. Her current research interests include Korean speech recognition, non-native speech recognition, and the system integration of an embedded system.

**Hong Kook Kim** received a B.S. degree in Control and Instrumentation Engineering from Seoul National University, Korea in 1988. He then received both M.S. and Ph.D. degrees in Electrical Engineering from the Korea Advanced Institute of Science and Technology (KAIST), Korea in 1990 and 1994, respectively. He was a senior researcher at the Samsung Advanced Institute of Technology (SAIT), Kiheung, Korea, from 1990 to 1998. During 1998–2003, he was a senior technical staff member with the Voice Enabled Services Research Lab at AT&T Labs-Research, Florham Park, NJ. Since August 2003, he has been with the School of Information and Communications, at the Gwangju Institute of Science and Technology (GIST) as a professor. His current research interests include speech recognition and coding, audio coding and 3D audio, and embedded algorithms and solutions for speech and audio processing for handheld devices.