6

Non-native Pronunciation Variation Modeling for Automatic Speech Recognition

Mina Kim[1], Yoo Rhee Oh[2] and Hong Kook Kim[2]
[1]*Mobile Communication Department, LG Electronics*
[2]*School of Information and Communications*
Gwangju Institute of Science and Technology
Korea

1. Introduction

Communication using speech is inherently natural, with this ability of communication unconsciously acquired in a step-by-step manner throughout life. In order to explore the benefits of speech communication in devices, there have been many research works performed over the past several decades. As a result, automatic speech recognition (ASR) systems have been deployed in a range of applications, including automatic reservation systems, dictation systems, navigation systems, etc.

Due to increasing globalization, the need for effective interlingual communication has also been growing. However, because of the fact that most people tend to speak foreign languages with variant or influent pronunciations, this has led to an increasing demand for the development of non-native ASR systems (Goronzy et al., 2001). In other words, a conventional ASR system is optimized with native speech; however, non-native speech has different characteristics from native speech. That is, non-native speech tends to reflect the pronunciations or syntactic characteristics of the mother tongue of the non-native speakers, as well as the wide range of fluencies among non-native speakers. Therefore, the performance of an ASR system evaluated using non-native speech tends to severely degrade when compared to that of native speech due to the mismatch between the native training data and the non-native test data (Compernolle, 2001). A simple way to improve the performance of an ASR system for non-native speech would be to train the ASR system using a non-native speech database, though in reality the number of non-native speech samples available for this task is not currently sufficient to train an ASR system. Thus, techniques for improving non-native ASR performance using only small amount of non-native speech are required.

There have been three major approaches for handling non-native speech for ASR: acoustic modeling, language modeling, and pronunciation modeling approaches. First, acoustic modeling approaches find pronunciation differences and transform and/or adapt acoustic models to include the effects of non-native speech (Gruhn et al., 2004; Morgan, 2004; Steidl et al., 2004). Second, language modeling approaches deal with the grammatical effects or speaking style of non-native speech (Bellegarda, 2001). Third, pronunciation modeling approaches derive pronunciation variant rules from non-native speech and apply the derived rules to pronunciation models for non-native speech (Amdal et al., 2000; Fosler-Lussier, 1999; Goronzy et al., 2004; Gruhn et al., 2004; Raux, 2004; Strik et al., 1999).

Source: Advances in Speech Recognition, Book edited by: Noam R. Shabtai,
ISBN 978-953-307-097-1, pp. 164, September 2010, Sciyo, Croatia, downloaded from SCIYO.COM

The remainder of this chapter is organized as follows. In Section 2, an overview of non-native speech recognition is investigated. After that, acoustic modeling, language modeling, and pronunciation modeling approaches are explained in Sections 3, 4, and 5, respectively. Then, a new pronunciation modeling method is proposed in Section 6 as a means of improving the performance of non-native speech recognition. In addition, the performance of a non-native ASR system adopting the proposed method is evaluated and compared to that employing conventional pronunciation model adaptation methods. Finally, we conclude our findings in Section 7.

2. Overview of non-native speech recognition

Recently, speech recognition technology has become more familiar in our lives (Goronzy et al., 2001), as numerous applications are increasingly adopting speech recognition systems. For example, voice dialing is possible based on either a user stating a name or a number, dictation systems are relatively common, and there are a number of voice-enabled automatic response systems available. However, when these ASR systems are used by non-native speakers, the performance of the system can rapidly degrade because of the mismatches between the native training data and the non-native test data (Compernolle, 2001).

Previously, several works have investigated the characteristics of non-native speech and the effect of non-native speech on ASR performance, some of which tried to explore the differences in characteristics between native and non-native speakers. For examples, the authors of (Sidaras et al., 2009) demonstrated that the duration and the first and second formant frequencies of English vowels spoken by Spanish speakers had different characteristics from those of native English speakers. Moreover, it was found that Spanish-accented English was perceived better when the listeners were trained with this form of English. Similarly, it was noticed that the tongue location of the English vowels by non-native speakers had different characteristics from that of native speakers (Wade et al., 2007). In addition, according to the work in (Alotaibi et al., 2010), unique consonants existed in some languages, such as four emphatic consonants of Arabic, and these unfamiliar consonants were found to be hard to perceive by non-native speakers. It was then found that when non-native speakers pronounced words containing these unfamiliar consonants, degradation of ASR performance could occur.

Other researchers have attempted to compare the ASR performance of both native and non-native speech. In (Wang et al., 2003), it was shown that the word error rate (WER) of an English ASR system by German speakers was 49.3% whereas that of native English speakers was 16.2%. Moreover, in (Steidl et al., 2004), an ASR system trained by German speakers provided WERs of 18.5% and 34.0% when tested by native German speakers and English speakers, respectively. However, when the same ASR system was trained by English speakers but tested by German speakers, the WER increased from 35.0% to 65.6%. Based on these previous works, it is evident that adjusting for different pronunciation characteristics between native and non-native speakers is crucial for improving the ASR performance of non-native speech.

In order to improve the ASR performance for non-native speech, we first need to prepare a non-native speech database to train the ASR system or adjust the system for non-native speech; then, each component of the ASR system can be adjusted for non-native speech. Depending on which ASR component is adapted or modified for non-native speech, we can classify the techniques developed for non-native speech as shown in Fig. 1. In brief, a typical

ASR system is composed of a front-end for extracting acoustic feature, acoustic models for representing recognition units with the acoustic features, a language model for covering language-specific grammar or syntax, and a pronunciation model for handling the phonology, phonotactics, or phonetics of the target language. Therefore, different techniques can deal with non-native ASR issues from acoustic modeling, language modeling, or pronunciation modeling points of view. In addition, it is also important to consider how to transform or compensate for acoustic features extracted from non-native speech into native speech. It is suggested here that to further improve ASR performance, a hybrid modeling approach can be used, one that combines some or all of the approaches mentioned above.

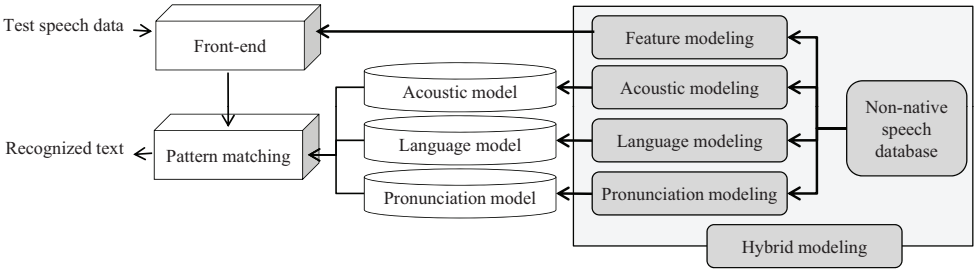

Fig. 1. Classification of techniques applied to non-native ASR.

1. Non-native speech database design
   In order to develop a non-native ASR system and investigate the characteristics of non-native speech, we first require non-native speech databases; Raab et al. (Raab et al., 2007) have previously reviewed such non-native speech databases.
2. Acoustic modeling approach
   Acoustic modeling approaches are used to adjust acoustic models and thereby improve the recognition performance of non-native speech (Gruhn et al., 2004; Morgan, 2004; Steidl et al., 2004). A simple way of adjusting acoustic models is to train them using a large amount of non-native speech. However, in practice it is rather difficult to collect a sufficient amount of non-native speech; therefore, acoustic models are usually adapted via a conventional acoustic model adaptation method, such as maximum likelihood linear regression (MLLR) and/or maximum a posteriori (MAP) methods (Yang et al., 2004). As an alternative, the acoustic models adjusted for non-native speech can also be obtained by interpolating the acoustic models for native speech and the acoustic models for the mother tongue (Steidl et al., 2004; Tan et al., 2007). In other words, the acoustic models trained with two different languages are combined to obtain the acoustic models for non-native speech. However, the most popular way of obtaining the adjusted acoustic models is to apply an adaptation technique with only small amount of adaptation data for non-native speech (Liu et al., 2008; Oh et al., 2007; 2009).
3. Language modeling approach
   Language modeling approaches deal with the grammatical effects or speaking styles of non-native speech, since non-native speakers tend to make a different sentence structure from native speakers (Bellegarda, 2001). However, there are relatively few research works in this area, compared to either the acoustic modeling approaches or the pronunciation modeling approaches (Huang et al., 2008; Raux et al., 2004; Steidl et al., 2004).

4. Pronunciation modeling approach
   Pronunciation modeling approaches first derive pronunciation variants from non-native speakers and then apply them to the pronunciation models for non-native speech. Usually, the variant pronunciations for each word are added to the pronunciation models, which is similar to a multiple pronunciation dictionary approach (Amdal et al., 2000; Fosler-Lussier, 1999; Goronzy et al., 2004; Gruhn et al., 2004; Raux, 2004; Strik et al., 1999).
   The pronunciation variants from non-native speakers can be derived by either knowledge-based or data-driven approaches (Strik et al., 1999). Note that knowledge-based approaches are based on linguistics or phonetic knowledge (Schaden, 2003; Tajchman et al., 1995; Wiseman et al., 1998), whereas data-driven approaches automatically derive pronunciation variants from non-native speech data and can be further classified into either a direct method (Amdal et al., 2000; Fosler-Lussier, 1999; Strik et al., 1999) or an indirect method (Amdal et al., 2000; Fosler-Lussier, 1999; Goronzy et al., 2004; Svendsen, 2004; Wolff et al., 2001).
   If many pronunciation variants are derived, the adapted pronunciation model becomes enlarged, resulting in performance degradation of the ASR system due to the fact that confusability in the pronunciation model is increased. Thus, several confusability reduction methods have also been proposed (Amdal et al., 2000; Hernandez-Abrego et al., 2004; Tsai et al., 2002).
5. Hybrid modeling approach
   Hybrid modeling approaches combine several modeling approaches, as described above, to further improve the performance of non-native ASR. In other words, acoustic or pronunciation modeling approaches can be combined in an MLLR and/or MAP adaptation framework (Goronzy et al., 2004; He et al., 2003; Liu et al., 2008; Oh et al., 2007; 2010; Tan et al., 2007). In particular, Bouselmi et al. (Bouselmi et al., 2007) proposed several combination schemes for pronunciation and MLLR/MAP acoustic model adaptations. On the other hand, pronunciation variant rules were decomposed into either pronunciation or acoustic variants (Oh et al., 2008). After that, pronunciation and acoustic model adaptations were applied to pronunciation and acoustic variants, respectively.
6. Feature-domain approach
   The feature-domain approach applies a feature adaptation method to compensate for mismatches between training and test conditions; the acoustic models are trained using native speech, but are tested using non-native speech. For example, Oh and Kim (Oh et al., 2010) applied a feature-space MLLR (fMLLR) adaptation with smoothing techniques to non-native ASR.

The next three sections will provide more detailed descriptions of the acoustic, language, and pronunciation modeling approaches.

3. Acoustic modeling approach

Because of the limited non-native speech database mentioned in Section 2, interpolating or adapting existing acoustic models using a small amount of non-native speech data is preferred, rather than attempting to train new acoustic models using large amounts of non-native speech data. Thus, in this section we introduce a number of acoustic modeling approaches in attempts to improve the performance of non-native ASR by using only a limited amount of non-native speech data.

As an effort to adapt acoustic models, several interpolation methods have been proposed, where two sets of acoustic models, the acoustic models trained with the target language and the acoustic models trained with the mother tongue of native speakers, are combined (Matsunaga et al., 2003; Steidl et al., 2004; Tan et al., 2007). Contrary to interpolating acoustic models, phone acoustic models of the target language were modified by adding an alternative path to the corresponding mother tongue phone acoustic models of non-native speakers (Bartkova et al., 2006; Bouselmi et al., 2006). Finally, the acoustic models were adapted by using non-native adaptation data via either an algorithm dedicated to non-native ASR or a conventional speaker adaptation method (Liu et al., 2008; Oh et al., 2007; 2009).

3.1 Retraining method

A retraining method generates non-native acoustic models by using a large amount of non-native speech data or retrains native acoustic models by using a moderately large sample of non-native speech data. These types of retraining methods are very simple but have several drawbacks, such as the following.

First, retraining methods require a large amount of non-native speech and their corresponding transcription data; however, these data are usually limited in quantity. Second, the transcriptions of a non-native speech database cannot be automatically generated since some non-native speech data contain various unpredictable pronunciations and structural errors. Third, the performance of ASR systems employing the retrained non-native acoustic models tends to drastically degrade for native speech (Oh et al., 2007).

For these reasons, several alternative methods have been proposed, which either interpolate the native and non-native acoustic models or adapt the native acoustic models based on a relatively small non-native database.

3.2 Interpolation method

In this subsection, we explain several interpolation methods, classified as either: 1) interpolation of native acoustic models of target language using non-native speech data (Steidl et al., 2004), and 2) interpolation of native acoustic models of target language based on native acoustic models of the mother tongue of non-native speakers (Tan et al., 2007).

3.2.1 Use of target language acoustic models

In this category, the acoustic model interpolation method is based on two assumptions. First, each non-native pronunciation has at least one similar native pronunciation in the target language, stemming from the fact that most languages have very similar phone inventories. Second, the native acoustic models of the target language are sufficient for adapting acoustic models for non-native speech.

The procedure of the acoustic model interpolation method is as follows:

Step 1. Generation of transcriptions based on native acoustic models

Each non-native utterance in a development set is recognized by the native acoustic models of the target language, which then automatically generates the transcriptions. According to the recognition results, each pronunciation in the lexicon is replaced by the recognized monophone such that highly specialized pronunciations in the lexicon are adapted.

Step 2. Selection of optimal interpolation partners

To select the optimal *K*-1 partners for acoustic model interpolation, each candidate partner is first interpolated based on the state of a hidden Markov model (HMM) of the target language, as shown in Eq. (1). Next, an *N*-best list of candidate partners is evaluated, and the first *K*-1 candidate partners are then selected from the *N*-best list.

Step 3. Interpolation of selected acoustic models

Since semi-continuous HMMs share the same set of output density probabilities, only the interpolation weights and the corresponding transition probabilities need to be adjusted in order to interpolate native acoustic models of the target language for non-native acoustic models. When there are *K*-1 interpolation partners for the state s_i of an HMM, the mixture weight $c_{i,m}$ of a state s_i of the HMM is adjusted as $\hat{c}_{i,m}$, based on the following equation:

$$\forall m \quad \hat{c}_{i,m} = \rho_1 \cdot c_{i_1,m} + \ldots + \rho_K \cdot c_{i_K,m} \tag{1}$$

where s_{i_1} represents s_i, and $s_{i_1},\cdots,s_{i_K}$ indicate the states of the corresponding interpolation partners of the state s_i. $\rightarrow c_{i_1}$ represents the mixture weight of s_i, and $c_{i_1},\cdots,c_{i_K}$ indicate the mixture weights of the states of the corresponding interpolation partners of the state s_i. In addition, ρ_1, …, ρ_K are the interpolation weights.

The interpolation weights indicate the probability from the original state s_i to the states of the corresponding interpolation partner and can be estimated using an expectation-maximization (EM) algorithm. After the interpolation weights are estimated, the corresponding transition probabilities can be determined in a similar manner.

3.2.2 Combined use of target language and mother tongue acoustic models

Tan and Besacier (Tan et al., 2007) proposed three interpolation methods based on the use of both the target language acoustic models and the mother tongue acoustic models of non-native speakers, which include 1) manual interpolation, 2) weighted least square based interpolation, and 3) eigenvoice based interpolation. The three acoustic model interpolation methods consist of two identical steps for preprocessing and one different step for the acoustic model interpolation.

Step 1. Investigation of phoneme mapping information

The mapping information on the phoneme substitutions for non-native speech is investigated using both the knowledge-based and the data-driven approaches.

- Knowledge-based approach
  Phoneme substitutions from the mother tongue of non-native speakers to the target language are first examined based on the international phonetic alphabet (IPA) tables (International Phonetic Association, 1999).
- Data-driven approach
  For a phoneme whose substitution information is not known from the IPA tables, a data-driven approach is applied using a phoneme confusion matrix. In other words, a forced alignment is first performed based on the target language acoustic models for each non-native utterance in a development set. Then, phoneme recognition is also performed using the mother tongue acoustic models for each non-native utterance. Next, the two phoneme sequences are aligned using time information in order to generate the phoneme confusion matrix. From the generated confusion matrix, the mapped phoneme having the highest probability is selected as the phoneme substitution for each phoneme.

Step 2. Regeneration of mother tongue acoustic models of non-native speakers

Before interpolating acoustic models, the mother tongue acoustic models of non-native speakers are reconstructed from the target language acoustic models in order to match the configuration of the target language acoustic models. For this task, the pronunciation dictionary of the mother tongue of non-native speakers is first modified using the investigated mapping information. The mother tongue acoustic models of non-native speakers are then reconstructed from the target language acoustic models by performing MLLR and MAP adaptations based on the speech corpus of the mother tongue of non-native speakers and the modified pronunciation dictionary.

- In the cases of manual and weighted least square based interpolations
  The mother tongue acoustic models of non-native speakers are reconstructed from the target language acoustic models by performing MLLR and MAP adaptations based on all the speech data of the mother tongue of non-native speakers and the modified pronunciation dictionary. In other words, speaker-independent acoustic models of the mother tongue are obtained as the mother tongue acoustic models.
- In the case of eigenvoice based interpolation
  For each native speaker of a speech training corpus, the target language acoustic models are reconstructed by performing MLLR and MAP adaptations using a subset of the speech corpus of the target language for the corresponding speaker and the original pronunciation dictionary. In other words, several sets of speaker-dependent acoustic models of the target language are obtained.
  Next, for each non-native speaker in a development speech corpus, the mother tongue acoustic models of non-native speakers are reconstructed from the target language acoustic models by performing MLLR and MAP adaptations using a subset of the speech corpus of the mother tongue for the corresponding speaker and the modified pronunciation dictionary. As a result, several sets of speaker-dependent acoustic models for the mother tongue are obtained.

Step 3.a. Manual interpolation of acoustic models

For the non-native acoustic models ($p_{interpolated}$) of a phoneme, the target language acoustic models ($p_{target\_language}$) for the phoneme are then interpolated based on the mother tongue acoustic models (p_{mother_tongue}) of the corresponding mapping phoneme, using the equation of

$$p_{interpolated} = w \cdot p_{target\_language} + (1 - w) \cdot p_{mother\_tongue} \tag{2}$$

where w ($0 \le w \le 1$) indicates an interpolation weight. In this method, the interpolation weight (w) is manually determined by experiments; this method is appropriate in the case that no non-native speech is available.

Step 3.b. Weighted least square based interpolation of acoustic models

If the non-native adaptation data are available, the interpolation weight can be predicted using the weighted least square. In other words, Eq. (2) can be rewritten as

$$A \cdot x = (p_{target\_language} \;\; p_{mother\_tongue}) \cdot \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = (p_{interpolated}) = b \tag{3}$$

where b is calculated as the speaker means obtained by a forced-alignment with the non-native adaptation data on the target language acoustic models.

Given A and b as in Eq. (3), the interpolation weight vector x can then be solved by using the weighted least square as

$$A^T \cdot \Sigma^{-1} \cdot A \cdot x = A^T \cdot \Sigma^{-1} \cdot b \tag{4}$$

where the speaker variance Σ is a weight since each mean does not have the same weight.

Step 3.c. Eigenvoice based interpolation of acoustic models

From all the generated sets of speaker-dependent acoustic models for the target language and the mother tongue, the means act as supervectors for creating a non-native space for eigenvoice based interpolation. Thus, a subset of these eigenvectors is selected for the interpolation.

3.3 Adaptation method

In order to compensate for mismatches between the native training data and the non-native test data of the target language, the native acoustic models of the target language are adapted using non-native speech such that the ASR performance for non-native speech can be improved. As a simple adaptation, traditional acoustic model adaptation methods, which are widely used for speaker adaptations or noise-robust ASR, can be applied. However, traditional MLLR and/or MAP adaptation methods adapt only speaker or environmental variability, not pronunciation variability from non-native speakers. Hence, this subsection focuses on acoustic model adaptation methods for handling pronunciation variability from non-native speakers (Oh et al., 2007; 2009).

3.3.1 Modified decision-tree based state-clustering method

The modified decision-tree based state-clustering method is performed in a decision-tree based state-tying step during construction of the acoustic models. The main procedure of the modified decision-tree based state-clustering method is as follows:

Step 1. Analysis of pronunciation variability of non-native speakers

Since the modified decision-tree based state-clustering method is based on the pronunciation variability of non-native speech, this pronunciation variability is first investigated in an indirect data-driven method that will be further explained in Section 6. In brief, for each utterance in a non-native development set, phoneme recognition is performed and then an N-best list of phoneme sequences is obtained. Next, the phoneme rule patterns that are derived from the recognized N-best lists are applied to a decision tree, C4.5 (Quinlan, 1993). As a result, the pronunciation variant rules are generated.

Step 2. Decomposition of pronunciation variability of non-native speakers

Among the derived pronunciation variant rules, *acoustic variants* are selected in the case that the default class ($phoneme_{default}$) of the pronunciation variant rule has a different phoneme from a target phoneme ($phoneme_{target}$). Other pronunciation variant rules are then determined as *pronunciation variants*. Note that only the acoustic variants are applied to the modified decision-tree based state-clustering method.

The acoustic and pronunciation variants can be briefly explained as follows:

- Acoustic variants, $phoneme_{variant_{acoustic}}$

  Acoustic variants are named since the pronunciation variant rules are applied in the acoustic modeling. In addition, it is assumed that the variants occurred due to

the different pronunciation characteristics between the target language and the non-native speaker's mother tongue. These acoustic variants can be placed in any context and thus they are also referred to as *context-independent variants*. For this reason, acoustic modeling is more appropriate than pronunciation modeling since pronunciation modeling adds variant pronunciations for each corresponding context and thereby increases the confusability.

- Pronunciation variants, $phoneme_{variant_{pronunciation}}$
  Pronunciation variants are named since the pronunciation variant rules are applied in the pronunciation modeling. In addition, it is assumed that the variants are due to the co-articulation effect. In the model, these pronunciation variants would be placed in a specific context with the left two phonemes and the right two phonemes, and thus they are also referred to as *context-dependent variants*. Consequently, pronunciation modeling can more properly handle the pronunciation variants by adding the corresponding variant pronunciations of each word.

Step 3. Adaptation in the state-tying step of acoustic model construction

The acoustic model adaptation is performed in the decision-tree based state-tying step of acoustic model construction using the acoustic variants. For a phoneme having no acoustic variants, a traditional state-tying step is applied, in which a decision tree for each target phoneme ($phoneme_{target}$) is utilized based on the states of the triphone acoustic models where the central phone of the triphone has the $phoneme_{target}$. However, for a phoneme having acoustic variants, a decision tree for each $phoneme_{target}$ is utilized by using the states of the triphone acoustic models in which the central phone of the triphone has either $phoneme_{target}$ or $phoneme_{variant_{acoustic}}$.

3.3.2 Modified MLLR adaptation method

A traditional MLLR adaptation method is commonly used for speaker or environment variants; however, the MLLR adaptation should be modified for non-native ASR (Oh et al., 2009). In other words, an MLLR/MAP adaptation for triphone models having pronunciation variations is performed to handle the pronunciation variability of non-native speakers. The main procedure of the modified MLLR adaptation method is as follows:

Step 1. Acquisition of pronunciation variations of non-native speech

The pronunciation variations of non-native speech are generated in an indirect data-driven approach, as will be explained in Section 6. Then, the only acoustic variants are selected by investigating the pronunciation variant rules in which the default class has a different phoneme as the target phoneme, as described in Section 3.3.1.

Step 2. Generation of regression classes

In this step, two separate sets of regression classes are generated; *overall regression classes* for the characteristics of non-native speakers or environments, and *pronunciation variation regression classes* for the pronunciation variations of non-native speech.

- For the overall regression classes
  All the acoustic models of the target language are pooled on the root node of a regression class tree and the overall regression classes are then generated by splitting the regression class tree to adapt the acoustic models of the target language for the characteristics of non-native speakers or environments.
- For the pronunciation variation regression classes
  Pronunciation variation regression classes are generated for each pronunciation having acoustic variants. That is, the acoustic models for both the target

pronunciation and the corresponding variant pronunciations are pooled on the root node of a regression class tree, and the pronunciation variation regression classes for the target pronunciation are then generated by splitting the regression class tree such that the acoustic models of the target language are adapted for the pronunciation variations of non-native speech.

In order to generate a regression class, the acoustic models pooled on the root node of a regression class tree are first split based on the criterion of the centroid splitting algorithm, using the Euclidean distance measure (Young et al., 2002). Then, each regression class is identified by using the acoustic models clustered on the leaf node of the regression class tree.

Step 3. Adaptation of acoustic models using MLLR and MAP adaptation methods

It is known that the combination of MLLR and MAP adaptations can further improve the ASR performance of non-native speech, as opposed to using either only the MLLR or MAP adaptations (Goronzy et al., 2004; He et al., 2003; Tan et al., 2007). Therefore, a second-pass adaptation method using both the MLLR and MAP adaptations is performed in order to adapt the acoustic models of the target language (Oh et al., 2009). In other words, for each regression class, the corresponding MLLR transformation matrix is first estimated via an EM algorithm based on the non-native adaptation data. Then, the adapted acoustic models are generated by applying a MAP adaptation with the non-native adaptation data and the estimated MLLR transformation matrix.

Step 4. Reconfiguration of the adapted acoustic models

Since one set of adapted acoustic models from the overall regression classes and several different sets of adapted acoustic models from the pronunciation variation regression classes are generated in Step 3) of this subsection, a single set of adapted acoustic models should be selected. To this end, for each pronunciation variation, the corresponding models in the adapted acoustic models from the overall regression classes are replaced by the acoustic models adapted by the corresponding pronunciation variation regression class. Accordingly, the reconfigured acoustic models can cover the characteristics of non-native speakers or environments as well as the pronunciation variations of non-native speech.

4. Language modeling approach

Language modeling approaches are associated with the different speaking styles or the grammatical effects of non-native speech. When compared to either the acoustic or pronunciation modeling approaches, there have been few research works reported on language modeling. Nevertheless, in this section, we explain the language modeling method for continuous word speech recognition and for pronunciation grammar (Huang et al., 2008; Raux et al., 2004; Steidl et al., 2004).

4.1 Interpolation with non-native language model

Non-native speakers tend to make different sentence structures from native speakers due to the syntactic characteristics of the mother tongue of non-native speakers. For handling such syntactic differences of non-native speech, Steidl et al. (Steidl et al., 2004) employed an adapted language model by combining the original native language model and the non-native language model. The non-native language model was generated by using the transliteration of a non-native speech database. In addition, Raux and Eskenazi (Raux et al.,

2004) generated a non-native language model for a language learning system having both native and non-native speech data. It was shown from subsequent experiments that both methods improved the recognition performance when compared to the native language model.

4.2 Unsupervised pronunciation grammar growing

Huang et al. (Huang et al., 2008) proposed an unsupervised pronunciation grammar growing method in order to obtain the grammar of the pronunciation variations of non-native speakers and to generate the pronunciation models for non-native speech. The method consisted of two steps: the construction of a pronunciation variation graph and the generation of the non-native grammar from the pronunciation variation graph.

The main procedure of the unsupervised pronunciation grammar growing method is as follows:

Step 1. Construction of a pronunciation variation graph

A pronunciation variation graph for a word starts with all the possible pronunciation variations including insertions, deletions, and substitutions. Thus, a huge search space is required for the pronunciation variation graph of a word. In the graph, a node indicates the possible pronunciation and an edge represents the possible transition between pronunciations. In order to reduce the search space of the graph, the possible pronunciations and transitions for a substitution are first constrained within the broad class information defined by linguistic experts. Next, the possible paths remaining for the pronunciation variations are evaluated by calculating the posterior probabilities of each phone pair (ph_{start}, ph_{end}) using the equation,

$$\frac{1}{N}\sum_{i}^{N} p(x_i \left| \lambda_{ph_{end}} \right.) = \frac{1}{N}\sum_{i}^{N} \frac{1}{\sqrt{(2\pi)^d \left| \Sigma_{ph_{end}} \right|}} \exp[-\frac{1}{2}(x_i - \mu_{ph_{end}})^T \Sigma_{ph_{end}}^{-1} (x_i - \mu_{ph_{end}})] \quad (5)$$

where x_i, N, and d indicate the i-th observation feature vector corresponding to ph_{start} in a training speech corpus, the number of observation feature vectors corresponding to ph_{start} in the training speech corpus, and the dimension of the observation feature vector, respectively. In addition, $\lambda_{ph_{end}}$, $\mu_{ph_{end}}$, and $\Sigma_{ph_{end}}$ represent the acoustic model, the mean vector, and the covariance matrix for the phone ph_{end}. In the experiment, paths that are greater than a predefined threshold remain in the pronunciation variation graph.

Next, the possible left-context and right-context dependent pronunciations are generated using both a target language pronunciation dictionary and a mother tongue pronunciation dictionary. Then, only the possible paths having context dependent pronunciations are extracted.

Step 2. Generation of non-native grammar

By using the constructed pronunciation variation graph, speech recognition is first performed and the pronunciation variation grammar is then optimized by removing the pronunciations that are incorrectly recognized or have unusual variants based on the recognition confidence and support score. Here, the word-level generalized posterior probability and the occurrence frequency of the pronunciation variation are used as the recognition confidence and the support score, respectively. The finally optimized pronunciation variation grammar is subsequently used to generate the multiple pronunciation dictionary for non-native speakers.

5. Pronunciation modeling approach

There are two approaches pertaining to pronunciation model adaptations for non-native speech: a knowledge-based approach and a data-driven approach (Strik et al., 1999). A knowledge-based approach uses pronunciation rules from phonological knowledge and develops a pronunciation dictionary based on the pronunciation rules. In the case of a data-driven approach, phonological rules for pronunciation adaptation are automatically generated from non-native speech and transcription data; as such, a subdivision into direct and indirect data-driven methods can be applied.

5.1 Knowledge-based method

In a knowledge-based method, phonologically obtained pronunciation rules are used to transform a baseform into a pronunciation variant. For example, the phonological rule

$$vowel + /b/ + /d/ \rightarrow vowel + /b/ + /D/ \tag{6}$$

is used to transform a consonant $/d/$ followed by a consonant $/b/$ into a fortis consonant $/D/$ in Korean. The phonological rules are derived based on linguistical and phonological knowledge according to known pronunciation variations of speech. Then, the phonological rules are applied to baseforms in a pronunciation dictionary.

As representatives of knowledge-based approaches, pronunciation rules from phonological knowledge were previously generated to develop a pronunciation dictionary based on pronunciation rules (Tajchman et al., 1995; Wiseman et al., 1998). Also, Schaden (Schaden, 2003) transformed canonical phonetic dictionaries of the target language into adapted dictionaries in order to model prototypical foreign-accented pronunciation variants.

5.2 Data-driven method

The primary advantage of the knowledge-based approach is that it can be applied to all corpora and especially to new words that are not introduced in the ASR system. However, a notable drawback of the approach is in that the rules are often very general, resulting in too many variants in the pronunciation dictionary, thereby increasing the confusability of pronunciation variations. Moreover, it should be noted that even if this approach is applied to an ASR system, it is unlikely that all aspects of non-native speech could be covered.

In order to compensate for such drawbacks of the knowledge-based approach, pronunciation variations are derived from speech signals in data-driven methods. Data-driven methods can be further classified into direct data-driven or indirect data-driven approaches. The direct data-driven approach derives pronunciation variants depending on pronunciation training databases, as proposed in (Amdal et al., 2000; Fosler-Lussier, 1999; Strik et al., 1999). When an ASR system employs the adapted pronunciation dictionary using a direct data-driven approach, some unseen words might appear during ASR testing. Thus, such a mismatch condition in the pronunciation model between ASR training and testing could degrade the performance of an ASR system.

On the other hand, an indirect data-driven method investigates pronunciation variability from the speech training data, derives the variant rules, and applies the variant rules in the ASR pronunciation dictionary to compensate for the variability (Amdal et al., 2000; Fosler-Lussier, 1999; Goronzy et al., 2004; Svendsen, 2004; Wolff et al., 2001). For example, pronunciation rules were derived using the speech training data, which in turn could be applied to generate one or

more baseforms of any vocabulary word in the pronunciation dictionary (Svendsen, 2004). In addition, variants were derived using a phoneme recognizer such that pronunciation rules could be constructed using a decision tree (Fosler-Lussier, 1999). 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 Verbmobil project reported in (Wolff et al., 2001). As another example, non-native speech was first examined using a phoneme recognizer to determine variants, and then variants caused by recognition errors were removed based on the statistics pertaining to the co-occurrences of phonemes (Amdal et al., 2000). In this way, Goronzy et al. (Goronzy et al., 2004) 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.

5.3 Confusability reduction of pronunciation dictionary

As described above, data-driven methods adapt a pronunciation dictionary after building variant rules from the derived pronunciation variants, whereas a knowledge-based method derives variant rules based on phonological and phonetic knowledge, and then adds alternatives of pronunciation variants into the pronunciation dictionary. However, the adapted pronunciation dictionary can have more than one element corresponding to a word in the pronunciation dictionary. Therefore, the system memory size must be increased in order to store the pronunciation dictionary, which also increases the computational complexity and results in a longer decoding time for ASR. It was also observed that adding pronunciation variants to the pronunciation dictionary increases the confusability, and that a large increase in confusability is probably one reason for only small improvements or even deteriorations of ASR performance (Tsai et al., 2002). By appropriately selecting the pronunciation variations, the confusability would be reduced. In order to mitigate this problem, several approaches have been previously reported, which will be discussed in Section 6.2.

6. Pronunciation model adaptation based on multiple pronunciation dictionary

In this section, we describe a new pronunciation model adaptation method and an optimization method of the adapted pronunciation models proposed in (Kim et al., 2008). In particular, Section 6.1 describes the proposed pronunciation adaptation method based on an indirect data-driven approach that adapts a pronunciation dictionary after building the variant rules from the derived pronunciation variants, resulting in a *multiple pronunciation dictionary*. This dictionary can have more than one element corresponding to a word in the pronunciation dictionary. Thus, a size optimization method of the multiple pronunciation dictionary is also proposed in Section 6.2, in which some confusable pronunciation variants in the pronunciation dictionary are removed. Finally, in Section 6.3, the performance of a non-native ASR system employing the proposed method is evaluated and compared with that using a conventional pronunciation model adaptation method.

6.1 Multiple pronunciation dictionary

Fig. 2 shows the main procedure of the proposed pronunciation variation modeling method based on an indirect data-driven approach that is applied to non-native speech. From the figure, the five steps of the procedure are as follows:

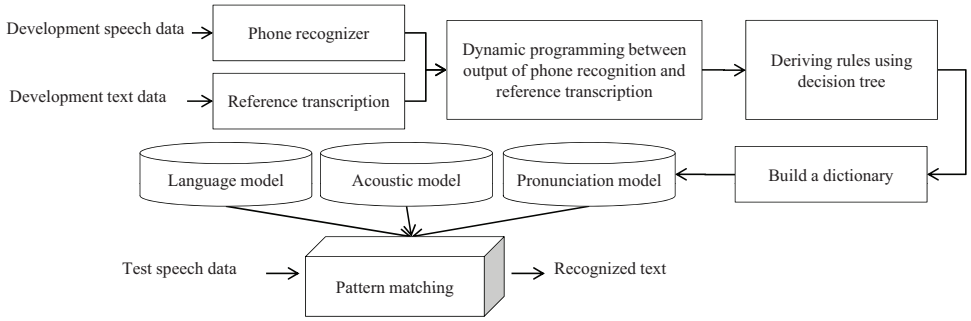

Fig. 2. Procedure of the proposed pronunciation variation modeling method based on an indirect data-driven approach applied to non-native ASR.

Step 1. Each utterance in a non-native development set is recognized using a phoneme recognizer.

Step 2. The recognized phoneme sequence is aligned using a dynamic programming algorithm based on the reference phoneme sequence transcribed by the native pronunciation dictionary, referred to as *reference transcription*.

Step 3. Using the alignment results of Step 2), variant phoneme patterns are obtained.

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

Step 5. Finally, pronunciation variations are generated from the pronunciation variation rules, allowing the pronunciation dictionary to be adapted for non-native ASR.

The details of each processing step are explained in further detail in the following subsections.

6.1.1 Phoneme recognition and aligned sequence

To derive the pronunciation rules, we first perform a phoneme recognition for each utterance in the non-native development set. As a result, we can obtain an N-best list of phoneme sequences for each utterance. However, there are no word boundaries in the list, which are required to differentiate inter-word pronunciation variations from cross-word pronunciation variations. To obtain these word boundaries, the recognized phoneme sequence is aligned on the basis of a dynamic programming algorithm and compared to the reference transcription with word boundaries.

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

$$L_2 - L_1 - X + R_1 + R_2 \rightarrow Y \tag{7}$$

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.

It is known from (Goronzy et al., 2004) that it is rather difficult to differentiate pronunciation variations from the substitution, deletion, and insertion errors incurred by phoneme recognition. Therefore, the recognition errors should be as small as possible; thus, three subsequent processes are applied to reduce these errors. First, we perform a Viterbi search based on the N-best lists. Second, we only extract a sentence or an isolated word included in

the development set if its phoneme recognition accuracy is over the predefined threshold. Third, if more than half of the neighboring phonemes of X in Eq. (7) are different from the neighboring phonemes of the target phoneme Y, this rule pattern is removed from the rule pattern set.

6.1.2 Decision-tree based rule derivation and pronunciation dictionary adaptation

Decision-tree based modeling is a popular method of deriving pronunciation variation rules (Fosler-Lussier, 1999; Wolff et al., 2001). Here, we use C4.5, a software extension of the basic ID3 algorithm designed by Quinlan (Quinlan, 1993). After the rule patterns are categorized by filtering errors, pronunciation variation rules are constructed by C4.5. Their attributes include the two left phonemes, L_1 and L_2, and the two right phonemes, R_1 and R_2, of the affected phoneme X. The output class is the target phoneme, where one decision tree is constructed for each phoneme. Next, each decision tree is converted into an equivalent set of the rules by tracing each path in the decision tree from the root node to each leaf node.
Next, a native pronunciation dictionary is adapted from these derived rules using C4.5, which results in a multiple pronunciation dictionary. For a more detailed description of adapting the pronunciation dictionary, refer to the work in (Kim et al., 2007).

6.2 Optimized multiple pronunciation dictionary

The size of the adapted multiple pronunciation dictionary could be much larger than that of the baseline pronunciation dictionary. As one solution to this problem, the confusability could be reduced by pruning the pronunciation variant rules based on either a rule probability, a rule probability using log likelihood, a decision tree, or another method. However, this approach does not take into account the interaction between words in a multiple pronunciation dictionary (Amdal et al., 2000). In other words, if a word is represented by several different phonetic sequences based on pronunciation variant rules and one of the sequences is similar to a phonetic sequence of another word, the confusability is further increased. Moreover, the confusability of words that have a smaller number of phonemes incurs errors in ASR systems (Hernandez-Abrego et al., 2004). Therefore, the number of phonemes in a word's sequence should be used as a measure of the confusability. In the following subsections, we propose a confusability measure and explain how the measure is applied to reduce the confusability in the multiple pronunciation dictionary of a non-native ASR system.

6.2.1 Confusability measure

Let M be a multiple pronunciation dictionary. It is assumed that the number of words in M is N_w and the i-th word, W_i, included in M, has $N_{p,i}$ pronunciation variants. Here, we denote $s_{i,j}$ as the j-th pronunciation variant belonging to the i-th word; i.e., $M=\{W_i \mid i = 1, \ldots, N_w\}$ and $W_i=\{s_{i,j} \mid j = 1, \ldots, N_{p,i}\}$. A confusability measure (CM) is then defined as

$$CM(s_{i,j}) = L(s_{i,j}) \cdot \min_{1 \le k \le N_w, k \ne i, 1 \le l \le N_{p,k}} \left\lfloor D(s_{i,j}, s_{k,l}) \cdot L(s_{k,l}) \right\rfloor \tag{8}$$

where $D(x,y)$ is the Levenshtein distance between x and y (Levenshtein, 1966). In addition, $L(x)$ is the number of phonemes of a pronunciation variant x, normalized by the maximal number of phonemes over all the pronunciations in M such that

$$l_{max} = max_{1 \le i \le N_w, 1 \le j \le N_{p,i}} \sharp(s_{i,j}) \quad (9)$$

where $\sharp(x)$ is defined as the number of phonemes in the pronunciation x. The goal of the proposed confusability measure defined in Eq. (8) is to detect pronunciation variants that are highly confusable so that ASR errors due to high similarities between the phonetic sequences of words in the multiple pronunciation dictionary can be reduced.
The normalized number of phonemes of a phonetic sequence x, is defined by

$$L(x) = l_x / l_{\max} \quad (10)$$

where $l_x = \sharp(x)$ and l_{max} is the maximum number of phonemes among all the sequences in M, as defined in Eq. (9). Eq. (10) contributes to the reduction of ASR errors because an ASR system tends to be more erroneous if the recognized word has a short phonetic sequence (Hernandez-Abrego et al., 2004). Therefore, the normalized number of a word's sequence can be used as a measure of the confusability.

6.2.2 Confusability reduction of multiple pronunciation dictionary

To reduce the confusability in the adapted multiple pronunciation dictionary, the confusability measure, defined in Eq. (8), for each pronunciation variant in the multiple pronunciation dictionary is first calculated. After that, all pronunciation variants except for the phonetic sequences obtained from the baseform are sorted according to their confusability measure scores. Finally, the pronunciation variants whose confusability measure scores are above a predefined threshold are used in constructing a pruned multiple pronunciation dictionary.

6.3 Speech recognition experiments

In order to evaluate the proposed pronunciation adaptation method, the baseline ASR system is first constructed. After that, we evaluate the performance of an ASR system using the pronunciation dictionary pruned by the proposed confusability reduction method, and compare it with that using the baseline pronunciation dictionary or the multiple pronunciation dictionary based on the indirect data-driven method.

6.3.1 Baseline ASR system

Especially, we want to develop a non-native ASR system that recognizes English spoken by Koreans. Thus, we need a training database spoken by native speakers to construct the baseline native ASR system. It is also required the native and non-native databases for developing and evaluating the non-native ASR system. In this subsection, we first describe the native and non-native databases. After that, we discuss how to construct each component of the baseline ASR system including ASR features, acoustic models, pronunciation and language models.

1. Training database
   As a training set for the baseline ASR system, we used a subset of the Wall Street Journal database (WSJ0) (Paul et al., 1992). The WSJ0 database was a 5000-word closed loop task for evaluating the performance of a large vocabulary continuous speech recognition (LVCSR) system. The training set consisted of 7,138 utterances recorded by a Sennheiser close-talking microphone and several far-field microphones, in which all utterances were sampled at a rate of 16 kHz.

2. Development and evaluation databases
   For developing and testing the proposed method, we used a subset of the Korean-Spoken English Corpus (K-SEC) (Rhee et al., 2004), comprised of English pronunciations spoken by both Korean and native English speakers. This database was divided into three parts: one was used for developing the pronunciation dictionary described in Section 6.1, and the others were evaluation subsets for both the baseline ASR system and an ASR system employing the proposed pronunciation modeling method. In other words, the two evaluation sets were comprised of utterances spoken by 49 Koreans and 7 native English speakers, respectively. The development set consisted of 11,125 isolated words spoken by 7 Koreans and 36 sentences by 98 Koreans, where each sentence had around 7 words. As a result, we had 7,299 isolated words and 3,123 continuous sentences for the development set. The two evaluation sets were made up of continuous sentences, in which each Korean or native speaker uttered 14 continuous sentences, resulting in a total of 146 words. In other words, we had 686 and 98 utterances for non-native speech and native speech, respectively.
3. Feature extraction
   For the baseline ASR system, we extracted 12 mel-frequency cepstral coefficients (MFCC) with logarithmic energy for every 10 ms analysis frame, and concatenated their first and second derivatives to obtain a 39-dimensional feature vector. During the training and testing, we applied a cepstral mean normalization to the feature vectors.
4. Acoustic models
   The acoustic models were based on 3-state left-to-right, context-dependent, 4-mixture, and cross-word triphone models, and they were trained using the HTK version 3.2 Toolkit (Young et al., 2002). All triphone models were expanded from 41 monophones, which included silence and pause models, and states of the triphone models were tied by employing a decision tree (Young et al., 1994). As a result, we had 9,655 physical triphones, 68,923 logical triphones, and 5,292 states, which was then referred to as the *baseline ASR system*.
5. Pronunciation and language models
   To develop a pronunciation dictionary, a back-off bigram language model was generated from a phoneme transcription of the training database, and the pronunciation dictionary was generated from a list of 41 phonemes with silence. In order to explore the behavior of pronunciation models based on the difference between the target language and the mother tongue, the pronunciation dictionary was only from the text of the test set. The pronunciation of each word was built from the CMU pronunciation dictionary (Weide, 1998) and any missing words from the CMU dictionary were transcribed manually. The pronunciation dictionary was comprised of 340 words, which was equal to the number of entries in the pronunciation dictionary. In addition, the relative ratio of the pronunciation dictionary size, defined as the average number of different pronunciations per word, was 1.

The performance of the baseline ASR system was tested using the two evaluation sets. Consequently, it was found that the average WERs of the baseline ASR system were 0.68% and 19.92% when the ASR system was tested by native speakers and by non-native speakers, respectively. This result confirmed the fact that performance of the ASR system tested by non-native speech could be exceedingly degraded.

Dictionary		WER (%)			Dictionary size (entry)	Relative ratio of dictionary	Real -time X
		Non-native	Native	Avg.			
a) Baseline		19.92	0.68	10.3	340	1	2.18
b) Multiple dictionary		21.96	0.68	11.32	512	1.51	3.5
c) Pruned multiple pronunciation dictionary							
Threshold	0	20.25	0.59	10.42	476	1.4	3.2
	0.1	**18.58**	**0.59**	**9.59**	**443**	**1.3**	**2.76**
	0.2	18.89	0.59	9.74	434	1.28	2.71
	0.3	18.93	0.59	9.76	428	1.26	2.67

Table 1. Performance comparison of an ASR system with a) the baseline pronunciation dictionary, b) a multiple pronunciation dictionary prior to reduction, and c) a pruned multiple pronunciation dictionary based on the proposed confusability reduction method. (Reprinted with permission from (Kim et al., 2008). Copyright IASTED/ACTA Press.)

6.3.2 Evaluation of the proposed pronunciation modeling method

To generate a multiple pronunciation dictionary, we performed a phoneme recognition and obtained a 200-best list for each utterance in the development set. As a phoneme recognizer, we used the baseline acoustic models, a phoneme based back-off bigram language model, and a pronunciation dictionary with a list of 41 phonemes with silence. By using the 200-best list, the performance of phoneme recognition was improved from 28.27% to 49.08%. In addition, the rule patterns could be generated using only phoneme sequences where the phoneme recognition accuracy was over 50%. After applying the rule patterns in C4.5, at a pruning option of 25%, we obtained 334 rules from the decision trees. Then, a multiple pronunciation dictionary was generated by adapting the baseline pronunciation dictionary from the obtained 334 rules. To reduce the confusability, we also applied the proposed optimization method to the adapted multiple pronunciation dictionary.

Table 1 compares the average WERs, the pronunciation dictionary size, and the ASR decoding time for the baseline ASR system and the ASR systems employing the multiple pronunciation dictionary and the pruned multiple pronunciation dictionaries according to pruning thresholds of 0, 0.1, 0.2, and 0.3. It can be seen in the table that the ASR system employing the multiple pronunciation dictionary increased the WER, compared to that employing the baseline pronunciation dictionary. The performance degradation incurred by the proposed multiple pronunciation dictionary was due to the increased confusability by improper pronunciation variants.

Next, the multiple pronunciation dictionary was pruned using the proposed confusability measure, and the average WERs of the ASR system using the differently pruned multiple pronunciation dictionaries are shown in the third row of Table 1. The table shows that the pruned multiple pronunciation dictionary constructed with a threshold of 0.1 gave the lowest average WER among all other dictionaries. That is, the average WERs of an ASR system using the pruned multiple pronunciation dictionary were 18.58% and 0.59% for non-native and native speech, respectively, which corresponded to relative WER reductions of 6.98% and 15.30%, compared to those of the baseline ASR system and an ASR system using the multiple pronunciation dictionary prior to pruning. Moreover, the ASR decoding time for the pruned multiple pronunciation dictionary was also reduced by 21.10% compared to that for the multiple pronunciation dictionary without pruning.

7. Conclusion

This chapter addressed issues associated with efficient pronunciation variation modeling for non-native automatic speech recognition (ASR), where non-native speech was mostly characterized by different pronunciations, speaking styles, and articulators of speakers from their native speech. The techniques for improving the performance of non-native ASR could then be classified into four approaches: acoustic modeling, language modeling, pronunciation modeling, and hybrid modeling approaches. We first reviewed these four approaches before proposing a new pronunciation model adaptation method.

In particular, the proposed pronunciation adaptation method was based on a multiple pronunciation dictionary, designed using an indirect data-driven method. However, this approach resulted in an increased search space for ASR decoding due to the increase of the pronunciation dictionary size. Therefore, a method for optimizing the size of the multiple pronunciation dictionary was also proposed, where a confusability measure based on the Levenshtein distance was introduced in order to remove some confusable pronunciation variants from the dictionary. To investigate the effects of the proposed approach on ASR performance, English was selected as the target language and English utterances spoken by Koreans were considered as the non-native speech. Subsequently, it was shown from the continuous non-native ASR experiments that an ASR system using the optimized multiple pronunciation dictionary could achieve an average word error rate reduction of 15.30%, with a relative reduction in computational complexity of 21.10%, compared to that achieved using the multiple pronunciation dictionary without optimization.

8. References

Alotaibi, Y. A. & Muhammad, G. (2010). Study on pharyngeal and uvular consonants in foreign accented Arabic for ASR, *Computer Speech and Language*, Vol. 24, No. 2, Apr. 2010, pp. 219-231.

Amdal, I.; Korkmazasky, F. & Suredan, A. C. (2000). Data-driven pronunciation modelling for non-native speakers using association strength between phones, *Proceedings of ISCA Tutorial and Research Workshop on Automatic Speech Recognition*, pp. 85-90, Paris, France, Sept. 2000.

Amdal, I.; Korkmazskiy, F. & Surendran, A. C. (2000). Joint pronunciation modelling of non-native speakers using data-driven methods, *Proceedings of 6th International Conference on Spoken Language Processing (ICSLP)*, pp. 622-625, Beijing, China, Oct. 2000.

Bartkova, K. & Jouvet, D. (2006). Using multilingual units for improved modeling of pronunciation variants, *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1037-1040, Toulouse, France, May 2006.

Bellegarda, J. (2001). An overview of statistical language model adaptation, *Proceedings of ITRW on Adaptation Methods for Speech Recognition*, pp. 165-174, Sophia Antipolis, France, Aug. 2001.

Bouselmi, G.; Fohr, D.; Illina, I. & Haton, J. P. (2006). Fully automated non-native speech recognition using confusion-based acoustic model integration and graphemic constraints, *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 345-348, Toulouse, France, May 2006.

Bouselmi, G.; Fohr, D. & Illina, I. (2007). Combined acoustic and pronunciation modelling for non-native speech recognition, *Proceedings of 8th Annual Conference of the International Speech Communication Association (Interspeech)*, pp. 1449-1452, Antwerp, Belgium, Aug. 2007.

Compernolle, D. V. (2001). Recognizing speech of goats, wolves, sheep and ... non-natives, *Speech Communication*, Vol. 35, Nos. 1-2, Aug. 2001, pp. 71-79.

Fosler-Lussier, E. (1999). Multi-level decision trees for static and dynamic pronunciation models, *Proceedings of 6th European Conference on Speech Communication and Technology (EUROSPEECH)*, pp. 463-466, Budapest, Hungary, Sept. 1999.

Goronzy, S.; Sahakyan, M. & Wokurek, W. (2001). Is non-native pronunciation modelling necessary?, *Proceedings of 7th European Conference on Speech Communication and Technology (EUROSPEECH)*, pp. 309-312, Aalborg, Denmark, Sept. 2001.

Goronzy, S.; Rapp, S. & Kompe, R. (2004). Generating non-native pronunciation variants for lexicon adaptation, *Speech Communication*, Vol. 42, No. 1, Jan. 2004, pp. 109-123.

Gruhn, R.; Markov, K. & Nakamura, S. (2004). A statistical lexicon for non-native speech recognition, *Proceedings of 8th International Conference on Spoken Language Processing (ICSLP)*, pp. 1497-1500, Jeju Island, Korea, Oct. 2004.

He, X. & Zhao, Y. (2003). Fast model selection based speaker adaptation for nonnative speech, *IEEE Transactions on Speech and Audio Processing*, Vol. 11, No. 4, July 2003, pp. 298-307.

Hernandez-Abrego, G.; Olorenshaw, L.; Tato, R. & Schaaf, T. (2004). Dictionary refinements based on phonetic consensus and non-uniform pronunciation reduction, *Proceedings of 8th International Conference on Spoken Language Processing (ICSLP)*, pp. 1697-1700, Jeju Island, Korea, Oct. 2004.

Huang, C.-L.; Wu, C.-H.; Li, H.; Hsieh, C.-H. & Ma, B. (2008). Unsupervised pronunciation grammar growing using knowledge-based and data-driven approaches, *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1097-1100, Hannover, Germany, June 2008.

International Phonetic Association. (2008). International Phonetic Association. *Handbook of the International Phonetic Alphabet: A Guide to the Use of the International Phonetic Alphabet*, Cambridge, UK.

Kim, M.; Oh, Y. R. & Kim, H. K. (2007). Non-native pronunciation variation modeling using an indirect data-driven method, *Proceedings of 10th biannual IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 231-236, Kyoto, Japan, Dec. 2007.

Kim, M.; Oh, Y. R. & Kim, H. K. (2008). Optimizing multiple pronunciation dictionary based on a confusability measure for non-native speech recognition, *Proceedings of Artificial Intelligence and Applications (AIA 2008)*, pp. 215-220, Innsbruck, Austria, Feb. 2008.

Levenshtein, V. I. (1996). Binary codes capable of correcting deletions, insertions, and reversals, *Soviet Physics Doklady*, Vol. 10, No. 8, Feb. 1996, pp. 707-710.

Liu, L.; Zheng, T. F. & Wu, W. (2008). State-dependent phoneme-based model merging for dialectal Chinese speech recognition, *Speech Communication*, Vol. 50, No. 7, July 2008, pp. 605-615.

Matsunaga, S.; Ogawa, A.; Yamaguchi, Y. & Imamura, A. (2003). Non-native English speech recognition using bilingual English lexicon and acoustic models, *Proceedings of IEEE*

International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 340-343, Hong Kong, China, Apr. 2003.

Morgan, J. (2004). Making a speech recognizer tolerate non-native speech through Gaussian mixture merging, *Proceedings of InSTIL/ICALL Symposium on Computer Assisted Learning*, paper 052, Venice, Italy, June 2004.

Oh, Y. R.; Yoon, J. S. & Kim, H. K. (2007). Acoustic model adaptation based on pronunciation variability analysis for non-native speech recognition, *Speech Communication*, Vol. 49, No. 1, Jan. 2007, pp. 59-70.

Oh, Y. R.; Kim, M. & Kim, H. K. (2008). Acoustic and pronunciation model adaptation for context-independent and context-dependent pronunciation variability of non-native speech, *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4281-4284, Las Vegas, NV, Apr. 2008.

Oh, Y. R. & Kim, H. K. (2009). MLLR/MAP adaptation using pronunciation variation for non-native speech recognition, *Proceedings of 11th biannual IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 216-221, Merano, Italy, Dec. 2009.

Oh, Y. R. & Kim, H. K. (2010). On the use of feature-space MLLR adaptation for non-native speech recognition, *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4314-4317, Dallas, TX, Mar. 2010.

Paul, D. & Baker, J. (1992). The design for the Wall Street Journal-based CSR corpus, *Proceedings of 5th DARPA Speech and Natural Language Workshop*, pp. 357-362, Harriman, NY, Feb. 1992.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers Inc., San Francisco, CA.

Raab, M.; Gruhn, R. & Noeth, E. (2007). Non-native speech databases, *Proceedings of 10th biannual IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 413-418, Kyoto, Japan, Dec. 2007.

Raux, A. (2004). Automated lexical adaptation and speaker clustering based on pronunciation habits for non-native speech recognition, *Proceedings of 8th International Conference on Spoken Language Processing (ICSLP)*, pp. 613-616, Jeju Island, Korea, Oct. 2004.

Raux, A. & Eskenazi, M. (2004). Using task-oriented spoken dialogue systems for language learning: Potential, practical applications and challenges, *Proceedings of InSTIL/ICALL Symposium on Computer Assisted Learning*, paper 035, Venice, Italy, June 2004.

Rhee, S.-C.; Lee, S.-H.; Kang, S.-K. & Lee. Y.-J. (2004). Design and construction of Korean-spoken English corpus, *Proceedings of 8th International Conference on Spoken Language Processing (ICSLP)*, pp. 2769-2772, Jeju Island, Korea, Oct. 2004.

Schaden, S. (2003). Generating non-native pronunciation lexicons by phonological rules, *Proceedings of International Congress of Phonetics Sciences*, pp. 2545-2548, Barcelona, Spain, Aug. 2003.

Sidaras, S. K.; Alexander, J. E. D. & Nygaard, L. C. (2009). Perceptual learning of systematic variation in Spanish-accented speech, *Journal of the Acoustical Society of America*, Vol. 125, No. 5. May 2009, pp. 3306-3316.

Strik, H. & Cucchiarini, C. (1999). Modeling pronunciation variation for ASR: A survey of the literature, *Speech Communication*, Vol. 29, Nos. 2-4, Nov. 1999, pp. 225-246.

Steidl, S.; Stemmer, G.; Hacker, C. & Noth, E. (2004). Adaptation in the pronunciation space for non-native speech recognition, *Proceedings of 8th International Conference on Spoken Language Processing (ICSLP)*, pp. 2901-2904, Jeju Island, Korea, Oct. 2004.

Svendsen, T. (2004). Pronunciation modeling for speech technology, *Proceedings of International Conference on Signal Processing and Communications (SPCOM)*, pp. 11-16, Bangalore, India, Dec. 2004.

Tajchman, G.; Fosler, E. & Jurafsdy, D. (1995). Building multiple pronunciation models for novel words using exploratory computational phonology, *Proceedings of 4th European Conference on Speech Communication and Technology (EUROSPEECH)*, pp. 2247-2250, Madrid, Spain, Sept. 1995.

Tan, T. & Besacier, L. (2007). Acoustic model interpolation for non-native speech recognition, *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1009-1012, Honolulu, HA, Apr. 2007.

Tsai, M.; Chou, F. & Lee, L. (2002). Improved pronunciation modeling by properly integrating better approaches for baseform generation, ranking and pruning, *Proceedings of ISCA Tutorial and Research Workshop (ITRW) on Pronunciation Modeling and Lexicon Adaptation for Spoken Language Technology*, pp. 77-82, Estes Park, CO, Sept. 2002.

Wade, T.; Jongman, A. & Sereno, J. (2007). Effects of acoustic variability in the perceptual learning of non-native-accented speech sounds, *Phonetica*, Vol. 64, Nos. 2-3, Aug. 2007, pp. 122-144.

Wang, Z.; Schultz, T. & Waibel, A. (2003). Comparison of acoustic model adaptation techniques on non-native speech, *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 540-543, Hong Kong, China, Apr. 2003.

Weide, H. (1998). *The CMU Pronunciation Dictionary, release 0.6*, Carnegie Mellon University, Pittsburgh, PA.

Wiseman, R. & Downey, S. (1998). Dynamic and static improvements to lexical baseforms, *Proceedings of ESCA Tutorial and Research Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, pp. 157-162, Rolduc, The Netherlands, May 1998.

Wolff, M.; Eichner, M. & Hoffmann, R. (2001). Automatic learning and optimization of pronunciation dictionaries, *Proceedings of ITRW on Adaptation Methods for Speech Recognition*, pp. 159-162, Sophia Antipolis, France, Aug. 2001.

Yang, J.; Pu, Y.; Wei, H. & Zhao, Z. (2004). Acoustic models adaptation in large vocabulary continuous Mandarin speech recognition for non-native speakers, *Proceedings of International Conference on Signal Processing (ICSP)*, pp. 687-690, Beijing, China, Sept. 2004.

Young, S.; Odell, J. & Woodland, P. (1994). Tree-based state tying for high accuracy acoustic modelling, *Proceedings of ARPA Human Language Technology Workshop*, pp. 307-312, Plainsboro, NJ, Mar. 1994.

Young, S.; Evermann, G.; Kershaw, D.; Moore, G.; Odell, J.; Ollason, D.; Povey, D.; Valtchev, V. & Woodland, P. (2002). *The HTK Book (for HTK Version 3.2)*, Cambridge, UK.

Advances in Speech Recognition

Edited by Noam Shabtai

ISBN 978-953-307-097-1
Hard cover, 164 pages
Publisher Sciyo
Published online 16, August, 2010
Published in print edition August, 2010

In the last decade, further applications of speech processing were developed, such as speaker recognition, human-machine interaction, non-English speech recognition, and non-native English speech recognition. This book addresses a few of these applications. Furthermore, major challenges that were typically ignored in previous speech recognition research, such as noise and reverberation, appear repeatedly in recent papers. I would like to sincerely thank the contributing authors, for their effort to bring their insights and perspectives on current open questions in speech recognition research.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Hong Kook Kim, Mina Kim and Yoo Rhee Oh (2010). Non-Native Pronunciation Variation Modeling for Automatic Speech Recognition, Advances in Speech Recognition, Noam Shabtai (Ed.), ISBN: 978-953-307-097-1, InTech, Available from: http://www.intechopen.com/books/advances-in-speech-recognition/non-native-pronunciation-variation-modeling-for-automatic-speech-recognition

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821