Chapter-2

Previous Work on Non-Native Speech Recognition

2.1 Speech Recognition for Non-Native Speakers

The advancement in communication has made cultural interactions between different parts of the world easier and more frequent. Although languages such as English, Chinese and some of other languages learned by people around the world at schools and universities, with the development of countries in Asia and other continents, more and more people around the world are embracing languages such as Mandarin, Indian, Arabic, Korean etc. Nowadays, apart from the native language, most people can speak at least one foreign language. Furthermore, people are more interested to travel foreign countries for vacation or business. They also often keep some common phrases with the help of the Internet and travel guides to make the communication easier with the locals.

Speech recognition technology has achieved tremendous advancement in the past decades. However, most of the works in speech recognition in the past focus on native speakers. Non-native speech as we see in previous section is different from native speech in term of phonology, pronunciation, vocabulary and grammars. These differences give rise to what is known as accent of a particular group of non-native speakers. What is the difference between non-native speech and dialects? For dialect speakers, there is no transfer of L1 (first language acquisition) like what happens for non-native speakers, because the dialect is often the first language of the speakers. However, variation from the ‘standard’ language can still happen in the areas of phonology, pronunciation, vocabulary and grammars. However, unlike non-native speech, dialect has commonly accepted phonology, pronunciation, vocabulary and grammar rules or standard among the speakers. Conversely, there is different degree of accent in non-native speech. The difficulty of non-native speech recognition is worsening by the number of languages available, and the limited amount of non-native resources. Three important components in speech recognition system are affected. They are the acoustic model, pronunciation model and
### 2.2 Non-native Modeling in Speech Recognition

Non-native speech has different characteristics compared to native speech. Hence, specific non-native models tailored to different non-native speaker groups have to be created to achieve better recognition speech performance. However, the lack of non-native resources implies that many of the conventional techniques proposed for native speakers are unable to be used effectively. Over the past decade, creative approaches have been developed for modeling non-native speech under the constraint of resources, by taking advantage of existing resources.

Automatic speech recognition system for non-native speakers has the same architecture as the conventional system. However, it may have an additional component which determines the accent of the speaker either manually or automatically. With this information, matching models which correspond to the accent of the speaker can be selected for decoding the speech.

### 2.3. General Adaptation Algorithms

General adaptation algorithms have proven to be effective for creating speaker specific model. By using a few utterances from a speaker, a speaker independent model can be adapted. Adaptation algorithms have also been used for adapting the environment conditions. The flexibility of adaptation algorithms, which are capable to work under limited resource constrains makes them an ideal choice to be employed for creating non-native models.

Two of the most popular adaptation algorithms in automatic speech recognition are Maximum Likelihood Linear Regression (MLLR) [9] and Maximum a Posteriori Estimation (MAP) [10, 11] found that adapting the target language acoustic model using MLLR or MAP with native speech of the speakers does not produce any improvement.
Contrary with this result, the acoustic models created from merging of the target language acoustic model with the target language acoustic model adapted with the native language of the speakers, have shown to be beneficial [12]. On the other hand, significant improvement can be obtained by adapting the target language acoustic model using small amount of non-native speech with.. MLLR or MAP. [14, 15] proposes to apply non-native speech with MAP adaptation and Polyphone Decision Trees Specialization (PDTS). PDTS [16] is a decision tree adaptation algorithm which is used to grow specialized non-native branches from a target language trees by pruning to the point where it can be inserted. The adapted tree represents contexts of the non-native speech data. Other general adaptation algorithm which has been tested on non-native speakers is [17]. It is an unsupervised speaker adaptation algorithm using incremental singular value decomposition (SVD) adaptation technique.

2.4 Pronunciation Modeling

Acoustic model defines elementary speech units using fine phonetic features which are related to mouth, tongue, vocal tract and others from speech. Pronunciation modeling on the other hand consists of creating the bigger word or syllable models using the acoustic units defined in acoustic model. Since in most cases, phoneme or phone is the acoustic unit employed in the acoustic model, a pronunciation dictionary (lexicon) can be built from a typical dictionary, since most of them contain descriptions of how words should be pronounced using International Phonetic Alphabet.[18]

If there is no description on the manner of pronunciation, then rules for converting the graphemes to phonemes have to be created. However, this requires an understanding of the language involved. An automatic grapheme to phoneme tool can be created for generating the 'standard' pronunciation models using linguistic rules. A manual verification of the generated pronunciation models is often required to correct words which are exception to the rules. In cases where rules for converting graphemes to phonemes do not exist, and there is limited understanding of the language involved, studies found that using
graphemes (context dependent) as the acoustic units for modeling pronunciation model can produce acceptable speech recognition performance, where it is only slightly worse compared to word modeled using phonemes [19, 20]. Note that, this also means that the grapheme units have to be trained in the acoustic model.

2.5 Pronunciation Dictionary

Typically a speech recognition system has a pronunciation dictionary which stores at least the base form representations or standard way for pronunciation of words or syllables. Hence, it is also natural to add the surface form or the variant pronunciation which maybe different from the base form into the pronunciation dictionary as another possible realization of the word [18].

One possible way to add pronunciation variants is through listening to the utterances, and writes down their pronunciations. However, this is time consuming and not necessarily produces better result than the automatic approach. A study shows that manual pronunciation modeling do not necessary outperforms automatic approach [21]. Automatic variants generation can be performed using data-driven approaches. The general procedure for finding pronunciation variants is by aligning the hypotheses obtained from non-native speech against the corresponding reference transcriptions to create phone confusion matrix. Pronunciation variants can be observed from the phone confusion matrix. The unobserved variants can be found by generalizing the variants found according to context by using decision trees, and optionally adding the variant probability from the decision trees for each word into the dictionary [22].

The procedure described above requires the usage of non-native speech. However, in many situations non-native speech is hard to acquire. [23] has attempted to generate pronunciation variants using the native phoneme set of the speaker. It is based on the hypothesis of cross-lingual transfer, where non-native speakers substitute target language phonemes with their native phonemes. The procedures are the same as described before for finding pronunciation variants using non-native speech. The only difference is that the target language speech is decoded by a phoneme recognition system of the source language (native language of the speaker). The phone confusions created from the
alignment are then used to create the decision trees. The pronunciation variants can then be subsequently retrieved from the trees. However, the results show that the new dictionary does not produce a significant improvement. Improvement is only obvious when the dictionary is used in conjunction with MLLR applied to the acoustic model with some speech from the speaker [18].

Different non-native speakers have different pronunciation habits which are specific to that group. [24] has proposed an automatic speaker clustering method for non-native speakers based on a list of manually defined vocalic substitutions. A vector is used to represent a dialogue session from a speaker. It contains the number of times a variant appears. Clustering is carried out using model-based k-means and the vectors are randomly assigned to one of the cluster initially. In the subsequent iteration, the vector is assigned to the cluster which gives the highest likelihood to the pronunciation variants observed. This step is executed until it converges. The pronunciation dictionary created for each cluster has shown to be able to reduce the WER of a speech recognition system for each group [18].