Computational Linguistics and Chinese Language Processing
Vol. 10, No. 3, September 2005, pp. 363-380

Modeling Pronunciation Variation for Bi-Lingual Mandarin/Taiwanese Speech Recognition

Dau-Cheng Lyu [*,], Ren-Yuan Lyu [*], Yuang-Chin Chiang[+] and Chun-Nan Hsu[**]**

Abstract

In this paper, a bi-lingual large vocaburary speech recognition experiment based on the idea of modeling pronunciation variations is described. The two languages under study are Mandarin Chinese and Taiwanese (Min-nan). These two languages are basically mutually unintelligible, and they have many words with the same Chinese characters and the same meanings, although they are pronounced differently. Observing the bi-lingual corpus, we found five types of pronunciation variations for Chinese characters. A one-pass, three-layer recognizer was developed that includes a combination of bi-lingual acoustic models, an integrated pronunciation model, and a tree-structure based searching net. The recognizer's performance was evaluated under three different pronunciation models. The results showed that the character error rate with integrated pronunciation models was better than that with pronunciation models, using either the knowledge-based or the data-driven approach. The relative frequency ratio was also used as a measure to choose the best number of pronunciation variations for each Chinese character. Finally, the best character error rates in Mandarin and Taiwanese testing sets were found to be 16.2% and 15.0%, respectively, when the average number of pronunciations for one Chinese character was 3.9.

Keywords: Bi-lingual, One-pass ASR, Pronunciation Modeling

1. Introduction

Words can be pronounced in more than one ways according to a lexicon; i.e., they usually have multiple pronunciations. Words are also pronounced differently by different people, a

[*] Chang Gung University, Taiwan
E-mail: rylyu@mail.cgu.edu.tw
[+] National Tsing Hua University, Taiwan
[**] Academia Sinica, Taiwan
E-mail: {daucheng, chunnan}@iis.sinica.edu.tw

[Received February 22, 2005; Revised July 22, 2005; Accepted August 15, 2005]

phenomenon called “pronunciation variation.” Pronunciation variation has been studied in the speech recognition field [Chen 1996; Cremelie 1996], and reports show that pronunciation variation can cause the performance of automatic speech recognizers to deteriorate if it is not well accounted for. A common approach to solving the pronunciation variation problem is to use pronunciation modeling; where multiple pronunciations are added to each lexeme in a lexicon in order to fit the acoustic data better.

A Chinese character is pronounced differently in different languages which use that Chinese character in their writing systems. The same character may or may not have the same meaning in such languages. For instance, the Chinese character “窗”(window) is pronounced “chuang^{11}” in Mandarin and “tang^{11}” in Taiwanese, and these are considered to be multiple pronunciations in a Mandarin/Taiwanese bi-lingual lexicon. “Chuang^{11}” is often mistakenly pronounced “cuang^{11}” (the un-retroflex of “chuang^{11}”) by native Taiwanese speakers, who do not have un-retroflex consonants in their language. This is a common cause of pronunciation variation. In the case of English, which has a more complex vowel inventory than the Han language family, the words “ear” and “year” are difficult for Mandarin speakers to tell apart. In other words, pronunciation variation is a natural and unavoidable phenomenon in a multi-lingual environment.

In this world of people who are well-connected by various types of communication devices, multi-lingual communication is necessary, and multi-lingual speech recognition is a must. This paper focuses on Mandarin-Taiwanese bi-lingual large vocabulary speech recognition, and the framework studied here is applicable to other language combinations as well.

Studies on the pronunciation variation problem have focused on two basic approaches, which are based on acoustic modeling or pronunciation modeling. For acoustic modeling, reports [Jurafsky *et al.* 2001] show that the triphone model can well capture variation resulting from phone substitution or phone reduction; other reports [Liu *et al.* 2003; Kam *et al.* 2003] show that well-trained triphone acoustic models can handle partial change of the pronunciation variation which depends on the context.

In pronunciation modeling, entries in the pronunciation dictionary include alternative pronunciation variations and associated probabilities, determined through either knowledge-based or data-driven approaches [Kipp *et al*. 1996; Zeppenfeld *et al*. 1997; Wiseman *et. al*. 1998; Wester 2003; Polzin *et al*. 1998; Peters *et al*. 1998; Bacchiani *et al*. 1999; Singh *et al*. 2002; Kessens 2003; Strik 2003.]. With the knowledge-based approach, variation information is obtained from research reports or pronunciation dictionaries. Techniques for obtaining the probabilities of possible pronunciation variations of a word in the data-driven approach include training decision trees, training an artificial network, using entropy, using the maximum likelihood criterion, and using the calculated phone confusion

matrix [Cremelie and Martens 1998; Riley *et al*. 1999; Kam *et al*. 2003; Fukada *et al*. 1997; 1998; Yang *et al.* 2000; Holter *et al*. 1999; Torre *et al*. 1997]. Techniques that achieve higher scores are chosen to serve as pronunciation variation rules.

In addition to the pronunciation variation within a word, substantial variation occurs across word boundaries [Finke *et al.* 1997; Fukada *et al.* 1998; Kessens *et al.* 1999.]. Due to the mono-syllabic nature of Mandarin and Taiwanese, pronunciation variation is complex, and we can identify five types of variation: (1) one orthography with pronunciation variation; (2) colloquial/literate switching; (3) tone sandhi; (4) one orthography with multiple pronunciations; (5) one pronunciation with multiple orthography. The first three types of variation occur in mono-lingual environment, while the last two occur in bi-lingual environments. Details will be given in Section 3.

The goal of this study was to construct a Mandarin/Taiwanese bi-lingual large vocabulary speech recognizer. We implemented a one-pass recognizer based on a bi-lingual acoustic model, an integrated pronunciation model, and a word searching net with tree-structured nodes. Most of the state-of-the-art speech recognizers, for either Western or Oriental languages, are implemented with the one-pass search strategy [Odell 1994; Aubert 1999; Hagen 2001]. In the acoustic modeling, one phonemic inventory called ForPA (Formosa Phonetic Alphabet) is used to transcribe bi-lingual corpora. [Lyu *et al.* 2004] According to this inventory, the acoustic models for similar sounds across languages are shared. In addition, we use an algorithm based on a decision tree to cluster similar acoustic models by means of the maximum likelihood criterion. In the pronunciation modeling, we integrate knowledge-based and data-driven approaches. If only the knowledge-based approach is adopted, some variation in the speech corpus can not be covered at all, while if only the data-driven approach is employed, the variation for each new corpus has to be determined. However, the more variations for each word there are in the searching net, the more the recognition time and confusability will increase. To limit the number of pronunciation variations for each Chinese character, we adopt a score based on the relative frequency ratio and choose the best average number of pronunciation variations. Furthermore, the tree-structured net directly uses each Chinese character as a searching node, which is also a new trial in the ASR field of Chinese languages.

This paper is organized as follows. Section 2 states the problem. Section 3 represents the proposed framework, which includes acoustic modeling, pronunciation modeling, and a searching net. In section 4, we report experimental results and analyze three different pronunciation models using a bi-lingual testing set. The final section is a summary.

2. Problem Statements

In recent decades, most of the speech recognition research related to the Chinese (also called the "漢" Han) language family has focused on Mandarin speech [Lee 1998; Liao *et al*. 2000]. Relatively few studies have focused on other languages [Lyu *et al.* 2000; Gau *et al.* 2000]. In this paper, we consider two languages in this language family, i.e., Mandarin and Taiwanese, simultaneously within the same framework of speech recognition. In Taiwan, Mandarin Chinese is the official language, and Taiwanese is the mother tongue of about three quarters of the population. Quite a few people speak Mandarin with an accent that is strongly influenced by Taiwanese, and when they speak Taiwanese, they mix in words from Mandarin. It appears that people in Southern China do much the same. If successful, we expect that this framework will work well for other combinations of Chinese languages.

In the Mandarin Chinese speech recognition system, a typical syllable decoder is implemented by searching a 3-layer network consisting of an acoustic model layer, a lexical layer, and a grammar layer, as shown in Figure 1. After the optimal syllable sequence or the syllable lattice is determined by the decoder, a syllable-to-character converter is applied to handle the homonym issue for the final text output, as shown in Figure 2. This framework works well and has long been used by the speech communication community. To generalize the system so as to incorporate more than one language, a straightforward approach is to extend the system with more acoustic models, more entries in the pronunciation dictionary, and more paths in the searching net. However, this will lead to the following difficulties:

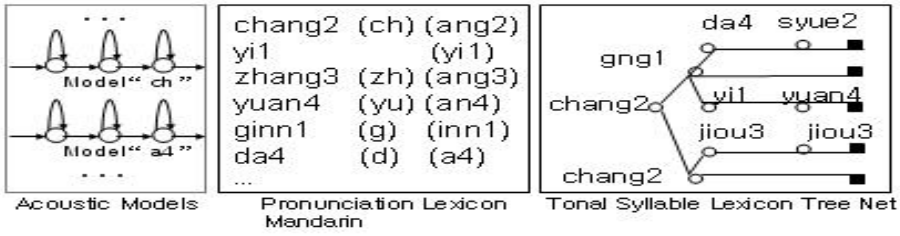

Figure 1. A 3-layer grammar searching net for syllable decoding

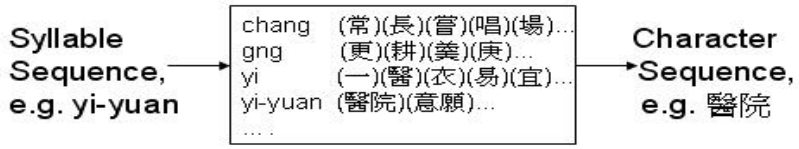

Figure 2. The syllable-to-character converter

1. In the case of multi-syllabic words such as "國家" (country), people rarely use Mandarin pronunciation for part of the word and Taiwanese pronunciation for the other part. It is, thus, impractical to generate all instances of all possible bi-lingual pronunciation variations of each

character in a word for a recognition network. Doing so will not only unnecessarily enlarge the searching space but also increase the time spent on decoding.

2. Generating multiple pronunciation lexicons efficiently is not a trivial task.

3. The language model for mixed languages is hard to estimate.

4. When new acoustic features like tones are added to the system, all 3 layers in syllable decoding and in the syllable-to-character converter should be modified. This also is not a trivial task.

3. Our Approach

Unlike some conventional approaches, which divide the recognition task into syllable decoding and character decoding, our proposed approach adopts a one-stage searching strategy, as shown in Figure 3, which decodes the acoustic feature sequence X directly to obtain the desired character sequence C*, no matter what languages are spoken. The decoding equation is, thus, as follows:

$$C^*(X) = \arg\max_{C} P(C \mid X). \tag{1}$$

In this framework, character decoding can be implemented by searching in a three-layer network composed of an acoustic model layer, a lexical layer, and a grammar layer, as shown in Figure 4. There are at least 2 critical differences between our framework and the conventional one. 1) In the lexicon layer, character-to-pronunciation mapping can easily incorporate multiple pronunciations caused by multiple languages, including Japanese, Korean, and even Vietnamese, which also use Chinese characters. 2) In the grammar layer, characters instead of syllables are used as nodes in the searching net. Under this ASR structure, we do not care which language the user speaks. No matter whether the language is Taiwanese, Mandarin or a mixture of them in one sentence, the ASR outputs the Chinese character only. This makes it language independent!

As in other multi-lingual researches [Young *et al*. 1997; Waibel 2000], determining how to efficiently and easily combine two languages in the acoustic and pronunciation models is very important. In the following two subsections, we will describe various approaches to integrating these two models in order to improve the recognition performance of ASR systems.

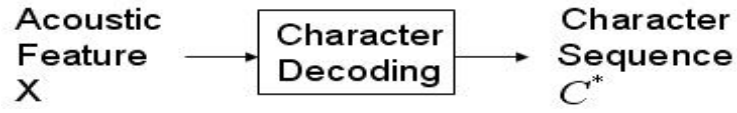

Figure 3. One-stage searching strategy for Chinese speech recognition

Figure 4. A unified 3-layer framework for multi-language Chinese speech recognition

3.1 Unified Bi-lingual Acoustic Modeling

It has been shown that the performance of acoustic models trained by combined speech database from multiple languages is better than that of models trained with speech data from a single language [Liu *et al.* 2003; Lyu *et al*. 2002]. For this reason, we use ForPA, which is an inventory of phoneme symbols, to transcribe the corpus of the two languages discussed here. Table 1 shows the statistical information of the phonemic inventory in different phonetic levels.

Table 1. The statistic information of all Mandarin (M) and Taiwanese (T) linguistic units in four levels: the numbers of Tonal Syllables (N_{TS}), Initials (N_I), Tonal Finals (N_{TF}), and context-dependent Initial/tonal Finals (N_{CDIF}). ∩and∪ mean intersection and union, respectively.

	M	T	M∪T	M∩T
N_{TS}	1288	2878	3519	647
N_I	17	19	22	14
N_{TF}	295	225	416	104
N_{CDIF}	1656	3496	4374	778

Sounds in different languages that are transcribed using the same phonemic symbols in ForPA share the same speech material. Combining two languages in this manner reduces the number of syllables by 21%. In order to easily integrate tone information, we used the context-dependent Initial and tonal Final as acoustic units, and trained these models by sharing the data which belonged to the same acoustic unit. Then, a divisive clustering algorithm was used to create context querying decision trees using four question sets, including an Initial set, a tonal Final set, the set of language properties, and a tonal information set. The above clustering approach could achieve significant improvement compared to previous results [Lyu *et al*. 2003].

Furthermore, in order to more efficiently merge the similar part of the sound for one phoneme or triphone model in both languages, we used a tying algorithm based on a decision tree to cluster the HMM models by using the maximum likelihood criterion [Liang *et al*. 1998; Lyu *et al*. 2002]. For the question sets, we used phonetic knowledge to design a total of 63 questions, including 10 language-dependent questions, 11 common questions, 28 Initial questions, and 14 Final questions. Then, the tree grew and split as we chose the optimal one among all the questions to maximize the increase in the likelihood scores or the decrease in uncertainty. Finally, the convergence condition was set to halt the growth of the decision tree. The acoustic model used in the experiment depended on the different splitting and convergence criteria adopted.

3.2 Pronunciation Modeling

The pronunciation model plays an important role in the Chinese character-based ASR engine [Liu *et al.* 2003; Huang *et al*. 2000]. It not only provides more choices during decoding if the speaker exhibits variations in pronunciation but also handles various speaking styles [Lyu *et al*. 2004]. As mentioned above, one Chinese character has more than two pronunciations in the combined phonetic inventory of Mandarin and Taiwanese. The factors of accent and regional migration can influence the pronunciation or speaking style of speakers too. Therefore, we identify the most common pronunciation variations in Taiwan in Table 2.

In Table 2, we list the five pronunciation variations that the Mandarin-Taiwanese bi-lingual recognizer can handle. Take the Chinese character "走" as an example. It is pronounced as "zau^{51}" in Taiwanese and means "to run" but is pronounced "zou^{21}" in Mandarin and means "to walk."

On the other hand, the total number of pronunciations in the pronunciation model for the decoding process is also important, because the more pronunciations are included in the lexicon, the more time the decoding process will take, and the less accurate of the ASR results will be [Strik *et al.* 1999]. The pronunciation variations will generate both improvements and deterioration in the ASR system, so previous research tried to find the optimal method to efficiently control the average pronunciation variations for one word in one language [Kesssens *et al*. 2003]. Our task is harder than that which deals with only one language. The reason is that one Chinese character must be mapped to at least two pronunciation variations, so cross-language confusion increases. In the following sections, we will propose two different methods, knowledge-based and data-driven methods, for obtaining rules of pronunciation variation.

Table 2. The five types of pronunciation variation rules in linguistic and phonological levels: 1. one orthography with pronunciation variations (OOPV); 2. colloquial literate switching (CLS) 3. tone sandhi (TS); 4. one orthography with multiple pronunciation (OOMP); 5. one pronunciation with multiple orthographies (OPMO). Other symbols and their meanings are: Chinese character (CC); Taiwanese or Mandarin pronunciations in literate style (TPL, MPL); Taiwanese or Mandarin Chinese character in colloquial style (TCC, MCC). The number [Yuen Ren Chao] following each syllable represents the tone patterns. e.g., zong^{51} means the syllable has a high-falling tone.

Within-language

(1)	CC	Base form	Surface form
OOPV	精彩	j**ing**^{55} cai^{21}	j**in**^{55} cai^{21}
	老師	lau^{21},**shii**^{55}	lau^{21} **sii**^{55}
	直的	d**it**^{55} e^{11}	d**i**^{55} e^{11}

(2)	MCC	MP	TCCL	TPL	TCCC	TPC
CLS	今天	Jin^{55}-ten^{55}	今天	gim^{33}-ten^{55}	今仔日	gin^{33}-na^{55}-lit^{55}
	明天	ming^{35}-ten^{55}	明天	bhing^{33}-ten^{55}	明仔載	mi^{33}-a^{55}-zai^{11}

(3)	CC	MP, isolated	MP, connected	TP, isolated	TP, connected
TS	總統府	zong**^{21}**,tong**^{21}** fu^{21}	zong**^{35}**-tong**^{35}**-fu^{21}	zong^{51},tong^{51} , hu^{51}	zong^{55}-tong^{55}-hu^{51}

Cross-language

(4)	CC	TP	MP
OOMP	走	zau^{51}	zou^{21}
	雨	u^{51},ho^{33}	yu^{21}
	行	giann^{15},hing^{15},hang^{15}	sing^{35},hang^{35}

(5)	Pronunciation	TCC	MCC
OPMO	jia^{55}-dan^{51}	［這裡］等	加蛋
	gau^{55}-gai^{51}	九［次］	高鈣

3.2.1 Knowledge-Based Method

As in [Wester *et al.* 2003], information about pronunciation can be derived from knowledge sources, such as pronunciation dictionaries hand-crafted by linguistic experts or extracted from the literature. In this approach, a pronunciation variation rule is simply the multiple pronunciations that appear in the lexicon for the same character. Associated probabilities can be calculated as follows. 1) the character-pronunciation pairs are derived; 2) the frequencies of the pairs are counted, and the relative frequency with respect to the total frequency of the

same Chinese character is calculated; 3) the pairs with high relative frequencies are kept as multiple pronunciation rules.

As our Mandarin knowledge source, we adopted the CKIP lexicon (http://ckip.iis.sinica.edu.tw/CKIP/) as our pronunciation lexicon source; it contains about 78,410 words. The length of one word in the lexicon varies from one Chinese character to ten, and the average of the length is 2.4 Chinese characters per word. As our Taiwanese knowledge resource, we adopted the Formosa lexicon (ForLex) [Lyu *et al*. 2000], which contains 104,179 words. The average length of one word in it is 2.8 Chinese characters. The pronunciation variation for each Chinese character was assigned a probability, which was estimated based on the frequency count of the pronunciations observed in both lexicons. The number of pronunciation variations for one Chinese character was 1.2 in the CKIP lexicon, and 2.1 in the Formosa lexicon. The number of pronunciation variations for Taiwanese was larger than that for Mandarin. The reasons are that most of the Chinese characters used in Taiwanese carry a classic literature pronunciation and a daily life pronunciation and that Taiwanese has much richer tone sandhi rules. Thus, the average number of pronunciation variations for one Chinese character is increased.

3.2.2 Data-Driven Approach

Although the regular pronunciation variations can be obtained from linguistic and phonological information, such as a dictionary, this information is not exhaustive; many phenomena in real speech have not yet been described. Therefore, another approach to deriving pronunciation variations from acoustic clues is presented below. All of the steps are also shown in Figure 5.

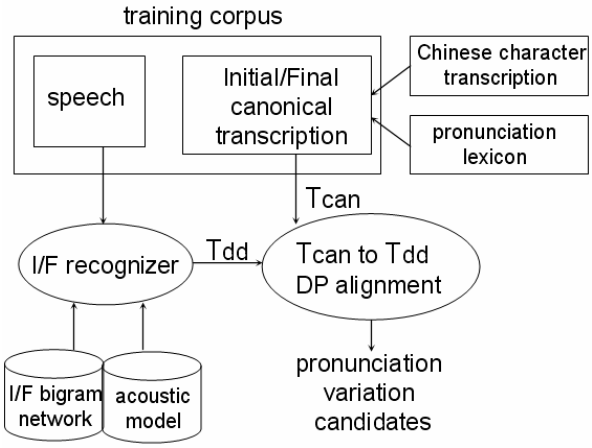

Figure 5. Diagram of pronunciation variations obtained with a data-driven approach.

First of all, the canonical transcription (T_{can}) is generated for each Chinese character in the phonetic levels of Initials and tonal Finals. Secondly, for each word in the utterance, a baseline recognition engine based on the Initial/tonal Final acoustic models is used to perform forced recognition, which adopts Viterbi search with an optional phonetic network [Strik 2003]. In this way, data-driven transcriptions (T_{dd}) of all the utterances in the training corpus can be obtained. Then, a dynamic programming algorithm is used to align T_{can} with T_{dd}. With this alignment, we can obtain a confusion table, which consists of pairs of easily confused phonetic units along with their likelihood scores.

A partial list of confusing phonetic units is shown as in Table 3. Using the above approach, we found that the major variation part in a syllable is Initial for both languages, especially in the retroflexion/un-retroflexion set. One of the possible reasons is that retroflex phonetic units exist only in Mandarin and most speakers usually do not accurately pronounce those retroflex units if their mother tongue is Taiwanese. These speakers tend to replace retroflex units with their un-retroflex counter parts.

Table 3. Some pronunciation variations obtained with the data-driven approach, where T_{can} and T_{dd} represent canonical transcription and data-driven transcription, respectively.

Mandarin				Taiwanese			
T_{can}	T_{dd}	T_{can}	T_{dd}	T_{can}	T_{dd}	T_{can}	T_{dd}
zh	z	s	sh	gh	g	p	t
sh	s	c	ch	g	d	r	l
ch	c	n	l	bh	l	h	t
z	zh	f	b	k	t	u3	u4

3.3 Searching Net

In the searching net, we use a large-vocabulary tree structured word net, because the perplexity can be reduced in the tree-structured searching net compare to the linear searching net. Figure 6 and Figure 7 show examples for a linear searching net and a tree-structured searching net, respectively. There were 5 words as searching paths in the linear net, and the equal probability of each path was set to be 1/5. We used equation 2 to calculate the entropy value based on the number of branches in each path, and we then used equation 3 to calculate the entropy from the perplexity. The perplexity of the linear searching net was found to be 5. This means that the perplexity in the linear searching net equals the number of distinct words. On the other hand, the procedure for determining the perplexity of the tree-structured searching net is described as follows. First, the Chinese characters are aligned according to their locations in multi-character words; characters that are in the same location in each word are considered to be redundant and, thus, eliminated. Finally, the entropy value is also

calculated based on the number of branches for each node, using equation (2). In the case shown in Figure 7, the entropy is 2.29, and the perplexity is 4.89, which is smaller than that of the linear searching net shown in Figure 6.

$$entropy = -\sum_i p_i \log_2 p_i \,, \tag{2}$$

$$perplexity = 2^{entropy} \,. \tag{3}$$

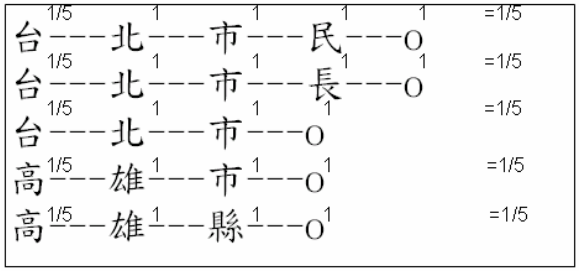

Figure 6. An example of an isolated linear searching net with its probability value.

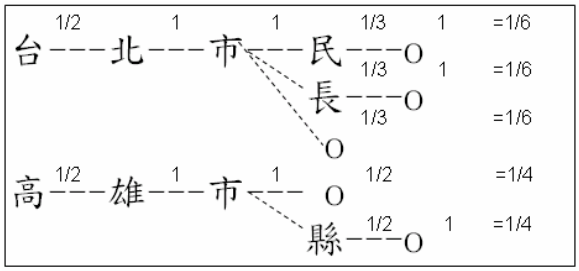

Figure 7. An example of an isolated tree-structured searching net with its probability value.

4. Experimental Results and Analysis

4.1 Corpus

All of the experiments employed a bi-lingual corpus, called ForSDa (Formosa Speech Database) [Lyu *et al.* 2004]. Both the training and testing data were read speech, which was recorded in the 16 kHz/16-bit wave-format in a normal office environment. The training set included a total of 89,164 utterances from 100 speakers, including 50 males and 50 females. Every speaker recorded speech in both languages. The utterances were phonetically balanced words, which were selected from a lexicon of about 40,000 words, using the phonetic abundant algorithm [Lyu 2003]. The length of the word varied from 1 to 6. The testing set included 2,000 utterances from 20 speakers; 10 speakers recorded speech in Taiwanese, and the other 10 speakers recorded speech in Mandarin. The statistics of the corpus employed here

are listed in Table 4.

Table 4. Statistics of the bi-lingual speech corpus used for training and testing sets. M: Mandarin, T: Taiwanese.

	Langue ID.	No. of Speakers	No. of Words	No. of Hours
Training	M	100	43078	11.3
	T	100	46086	11.2
Test_M	M	10	1000	0.28
Test_T	T	10	1000	0.28

4.2 Experimental Setup

The experiment setup can be described as follows. Firstly, we used context dependent Initials and tonal Finals with 16 Gaussian mixtures in HMM modeling. The feature vectors used in the HMM included 42 components, with 12 mel-frequency cepstral coefficients (MFCCs), normalized log energy, and pitch with their first and second order derivatives. Secondly, in pronunciation modeling, we used three models, which included knowledge-based, data-driven, and combined approaches, called P_{KW}, P_{DD} and P_{KW+DD}, respectively. The average number of pronunciations for one Chinese character for each pronunciation lexicon was 3.2, 2.7 and 3.9 for P_{KW}, P_{DD} and P_{KW+DD}, respectively. Finally, the tree-structured searching net consisted of 30,000 words, and the word perplexity of the net was 15,249. This means that there where almost 15,249 candidates for each input speech utterance in the decoding phase. Additionally, the output of the recognizer was Chinese characters; therefore, we evaluated the performance based on the Chinese character error rate (CER).

4.3 Experiment Results

Table 5 shows the CER results for pronunciation modeling with the Taiwanese and Mandarin testing sets. We can draw two conclusions; firstly, when the pronunciation model P_{KW+DD} was used, the CER was minimal for both languages. The reason is that P_{KW+DD} could capture both within-language and cross-language pronunciation variations. Secondly, the CER of the Test_M set with P_{DD} was better than that with P_{KW}, but the CER of the Test_T set was worse. A possible reason is that most of the pronunciation variations in Taiwanese can be found in the dictionary or lexicon source, such as tone sandhi or colloquial/literate switching. However, in Mandarin, most of the pronunciation variations are due to co-articulation, regional accents, speaking rates, speaking styles, etc. Such types of the variation can only be captured in speech data, not in lexicons. Therefore, the CER of Test_M dropped about 2.2% (17.9%-20.1%) when P_{DD} was used compared to the result obtained with P_{KW}, but the CER of Test_T increased 0.7% (18.3%-17.6%).

Table 5. CER (Character Error Rate) results for three pronunciation models with two testing sets. P_{KW}: pronunciation modeling using the knowledge-based method; P_{DD}: pronunciation modeling using the data-driven approach; P_{KW+DD}: pronunciation modeling using both P_{KW} and P_{DD}.

	P_{KW}	P_{DD}	P_{KW+DD}
Test_M	20.1%	17.9%	16.2%
Test_T	17.6%	18.3%	15.0%

4.4 Error Analysis

The addition of pronunciation variants to a lexicon increases the confusability, especially if the lexicon is large. Here, the large increase in confusability was probably the reason why only a small improvement or even deterioration in performance is found. The experimental results represented in Figure 8 show the CER performance as a function of the number of pronunciation variations for each Chinese character. It can be seen that the CER decreased when the average number of pronunciation variations increased. The lowest CER results were obtained when the number of pronunciation averaged 3.9. This was achieved using P_{KW+DD} and by eliminating variants with probabilities smaller than 0.1.

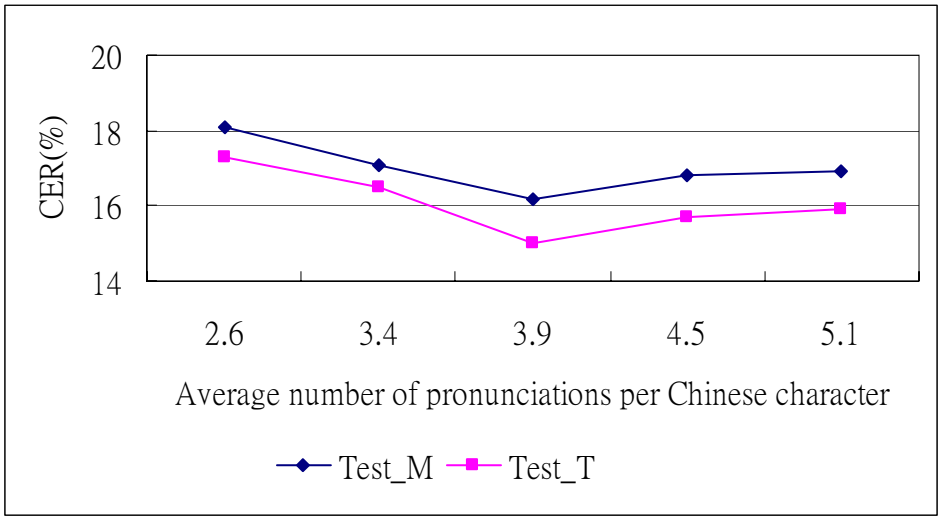

Figure 8. CER performance for P_{KW+DD} with different numbers of pronunciation variations per Chinese character.

Moreover, the error types mentioned above can be classified into the following 3 sets.

A. Cross-language homophonic confusion

This kind of error is just like the fifth term in Table 1, and occurs when different Chinese words belonging to different languages have the same or similar pronunciation. Therefore, the

confusion of choosing the final Chinese words will occur during the decoding phase. For example, the pronunciation of the Chinese word "星系" in Mandarin, that is, /sing^{55}-si^{51}/, is similar with that of "先死" in Taiwanese, that is, /sing^{33}-si^{51}/. The same is true of "高等" in Mandarin, pronounced /gau^{55}-dng^{13}/, and "教堂" in Taiwanese, pronounced, /gau^{51}-dng^{13}/.

B. Within-language homophonic confusion

This type of error is similar to the first error type, but it only occurs within one language. For example, the Chinese words "穢亂" and "會亂" have the same pronunciation, that is, /huei^{51}-luan^{51}/, in Mandarin, and "交待" and "交代" both have the same pronunciation in Mandarin, that is, /jiau^{55}-dai^{51}/, and in Taiwanese, that is, /gau^{55}-dai^{55}/.

C. Tone confusion:

This kind of error occurs due to mismatch between the tone pattern and speech features. We add the tone vectors to the feature parameters, the words, "水餃" and "睡覺" can be easily discriminated a tonal phase. However, there is also a side effect if the acoustic model in the tone aspect is not robust enough. A major tone error may be due to confusion between a high-level (55) tone and a mid-level (35) tone. Another major error may due to the confusion between a mid-falling (31) tone and a high-falling tone. Following are some tone confusion examples:

1) "縫補" /fng^{35}-bu^{31}/ and "蜂舞" /fng^{55}-u^{31}/.
2) "股票" /gu^{31}-piau^{51}/ and "顧票" /gu^{51}-piau^{51}/.

Most of the performance deterioration observed in this experiment was caused by the above error types; however, the performances of deterioration are smaller than that of improvements by adding pronunciation variations to the lexicon. Therefore, finally, we got an improvement in CER result.

5. Conclusion

As mentioned in the introduction, the goal of this study was to convert both Taiwanese and Mandarin speech into Chinese characters. In order to deal with the issues of multiple pronunciations and pronunciation variations for each Chinese character in these two languages in the ASR system, we developed a one-pass, three-layer recognizer, which includes combined bi-lingual acoustic models, an integrated pronunciation model and a tree-structure-based searching net. In the pronunciation model, an integrated method is used to combine the knowledge-based and data-driven approaches. Since the knowledge-based approach is used, homophony in Chinese characters can be addressed, and since the data-driven approach is employed, speakers' accents or styles can also be dealt with.

The experimental results showed that the CER could be improved by using the three different pronunciation models. The best performance was 16.2% and 15.0% for the testing sets Test_M and Test_T, respectively, where the perplexity was 15,249 for 30,000 words, and the P_{KW+DD} pronunciation model was used. In addition, in order to limit the side effect where in the increase in the size of the pronunciation lexicon causes the performance to deteriorate, the average number of pronunciations for both languages was 3.9.

The method proposed in this paper has been applied to two languages in the Chinese language family, but it can be easily extended to other languages or dialects. We have also discussed the major five pronunciation variations found in Taiwan. This is the first work, to the best of our knowledge, that has systemically investigated pronunciation variations in Mandarin and Taiwanese speech conversion to Chinese characters using ASR technology.

References

Aubert, X., "One pass cross word decoding for large vocabularies based on a lexical tree search organization," In *Proceedings of the European Conference on Speech Communication and Technology*, 1999, Budapest, Hungary, pp. 1559-1562.

Bacchiani, M., and M. Ostendorf, "Joint lexicon, acoustic unit inventory and model design," *International Journal of Speech Communication*, 29(2-4), 1999, pp. 99-114.

Chao, Y. R., Tone contour, http://en.wikipedia.org/wiki/Tone_contour/, 1979.

Cremelie, N., and J.-P. Martens, "In search of pronunciation rules," In *Proceedings of the European Speech Communication Association (ESCA) Workshop on Modeling Pronunciation Variation for Acoustic Speech Recognition*, 1998, Rolduc, Kerkrade, pp. 103-108.

Downey, S., and R. Wiseman, "Dynamic and static improvements to lexical baseforms," In *Proceedings of the Workshop on Modeling Pronunciation Variations*, 1998, Roldue, pp. 157-162.

Finke, M., and A. Waibel, "Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition," In *Proceedings of the European Conference on Speech Communication and Technology*, 1997, Rhodos, Greece, pp. 2379-2382.

Fukada, T., and Y. Sagisaka, "Automatic generation of a pronunciation dictionary based on a pronunciation network," In *Proceedings of the European Conference on Speech Communication and Technology*, 1997, Rhodos, pp. 2471-2474.

Fukada, T., T. Yoshimura, and Y. Sagisaka, "Automatic generation of multiple pronunciations based on neural networks and language statistics," In *Proceedings of the European Speech Communication Association (ESCA) Workshop on Modeling Pronunciation Variation for Acoustic Speech Recognition*, 1998, Rolduc, Kerkrade, pp. 103-108.

Holter, T., and T. Svendsen, "Maximum likelihood modelling of pronunciation variation," *International Journal of Speech Communication*, 29, 1999, pp. 177-191.

Huang, C., E. Chang, J.L. Zhou, and K.F. Lee, "Accent Modeling Based on Pronunciation Dictionary Adaptation for Large Vocabulary Mandarin Speech recognition," In *Proceedings of the International Conference on Spoken Language Processing*, 2000, Beijing.

Jurafsky, D., W. Ward, J. Zhang, K. Herold, X. Yu, and S. Zhang, "What kind of pronunciation variation is hard for triphones to model?" In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 2001, Salt Lake City, Utah, pp. 577-580.

Kam, P., and T. Lee, "Modeling pronunciation variation for Cantonese speech recognition," In *Proceedings of ISCA ITR-Workshop on Pronunciation Modeling and Lexicon Adaptation*, 2002, Colorado, USA, pp.12-17.

Kam, P., T. Lee, and F. Soong, "Modeling Cantonese pronunciation variation by acoustic model refinement," In *Proceedings of the 8th European Conference on Speech Communication and Technology*, 2003, Geneva, Switzerland, pp.1477-1480.

Kessens, J.M., C. Cucchiarini, and H. Strik, "A data-driven method for modeling pronunciation variation," *International Journal of Speech Communication*, 40, 2003, pp. 517-534.

Kessens, J.M., H. Strik, and C. Cucchiarini, "Modeling pronunciation variation for ASR: Comparing criteria for rule selection," In *Proceedings of the Workshop on Pronunciation Modeling and Lexicon Adaptation*, 2002, Estes Park, USA, pp. 18-23.

Kessens, J.M., M. Wester, and H. Strik, "Improving the Performance of a Dutch CSR by Modeling Within-word and Cross-word Pronunciation Variation," *International Journal of Speech Communication on Special issue of 'Modeling Pronunciation Variation for Automatic Speech Recognition'*, 29(2-4), 1999, pp. 193-207.

Kipp, A., M.-B. Wesenick, and F. Schiel, "Automatic detection and segmentation of pronunciation variants in German speech corpora," In *Proceedings of the International Conference on Spoken Language Processing*, 1996, Philadelphia, USA, pp. 106-109.

Lee, T., W. Lau, Y. W. Wong, and P.C. Ching, "Using tone Information In Cantonese Continuous Speech Recognition," *ACM Transactions on Asian Language Information Processing*, 1, 2002, pp. 83-102.

Liang, M.S., R.Y. Lyu, and Y.C. Chiang, "An efficient algorithm to select phonetically balanced scripts for constructing corpus," In *Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering,* 2003, Beijing, China.

Liang, P.Y. , J. L. Shen, and L. S. Lee, "Decision Tree Clustering for Acoustic Modeling in Speaker-Independent Mandarin Telephone Speech Recognition," In *Proceedings of the International Symposium on Chinese Spoken Language Processing* , 1998, Singapore, pp. 207-211.

Liao, Y. F., N. Wang, M. Huang, H. Huang, and F. Seide, "Improvements of the Philips 2000 Taiwan Mandarin Benchmark System," In *Proceedings of the International Conference on Spoken Language Processing*, 2000, Beijing. pp. 298-301.

Liu, Y., and P. Fung, "Modeling partial pronuncia-tion variations for spontaneous Mandarin speech recog-nition," *International Journal of Computer Speech and Language*, 17, 2003, pp. 357-379.

Liu, Y., and P. Fung, "Partial change accent models for accented Mandarin speech recognition," In *Proceedings of the IEEE Workshop on ASRU*, 2003, St. Thomas, U.S. Virgin Islands.

Liu, Y., and P. Fung, "State-Dependent Phonetic Tied Mixtures with Pronunciation Modeling for Spontaneous Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, 12, 2004, pp. 351-364.

Lyu, D.C., B.H. Yang, M.S. Liang, R.Y. Lyu, and C.N. Hsu, "Speaker Independent Acoustic Modeling for Large Vocabulary Bi-lingual Taiwanese/Mandarin Continuous Speech Recognition," *In Proceedings of the 9th Australian International Conference on Speech Science & Technology*, 2002, Melbourne, Australia.

Lyu, D.C., M.S. Liang, Y.C. Chiang, C.N. Hsu, and R.Y. Lyu, "Large Vocabulary Taiwanese (Min-nan) Speech Recognition Using Tone Features and Statistical Pronunciation Modeling," In *Proceedings of the 8th European Conference on Speech Communication and Technology*, 2003, Geneva, Switzerland.

Lyu, D.C., M.S. Liang, Y.C. Chiang, C.N. Hsu and R.Y. Lyu, "Large Vocabulary Taiwanese (Min-nan) Speech Recognition Using Tone Features and Statistical Pronunciation Modeling," In *Proceedings of the European Conference on Speech Communication and Technology*, 2003, Geneva, Switzerland.

Lyu, R.Y., C.Y. Chen, Y.C. Chiang, and M.S. Liang, "Bi-lingual Manda-rin/Taiwanese(Min-nan), Large Vocabulary, Continuous Speech Recognition System Based on the Yong-yong Phonetic Alphabet," *In Proceedings of the International Conference on Spoken Language Processing*, 2000, Beijing, China.

Lyu, R.Y., D.C. Lyu, M.S. Liang, M.H. Wang, Y.C. Chiang, and C.N. Hsu, "A Unified Framework for Large Vocabulary Speech Recognition of Mutually Unintelligible Chinese "Regionalects"," In *Proceedings of the 8th International Conference on Spoken Language Processing*, 2004, Jeju Island, Korea.

Lyu, R.Y., M.S. Liang, and Y.C. Chiang, "Toward Constructing A Multilingual Speech Corpus for Taiwanese (Minnan), Hakka, and Mandarin," *International Journal of Computational Linguistics and Chinese Language Processing*, 9(2), 2004, pp. 1-12.

Odell, J.J., V. Valtchev, P.C.Woodland, and S.J.Young, "A One Pass Decoder Design for Large Vocabulary Recognition," In *Proceedings of Human Language Technology Workshop*, 1994, pp. 405-410.

Peters, S.D., and P. Stubley, "Visualizing speech trajectories," In *Proceedings of the European Speech Communication Association (ESCA) Workshop on Modeling Pronunciation Variation for Acoustic Speech Recognition*, 1998, Rolduc, Kerkrade, pp. 103-108.

Polzin, T.S., and A.H. Waibel, "Pronunciation variations in emotional speech," In *Proceedings of the European Speech Communication Association (ESCA) Workshop on*

Modeling Pronunciation Variation for Acoustic Speech Recognition, 1998, Rolduc, Kerkrade, pp. 103-108.

Riley, M., W. Byrne, M. Finke, S. Khudanpur, A. Ljolje, J. McDonough, H. Nock, M. Saraclar, C. Wooters, and G. Zavaliagkos, "Stochastic pronunciation modelling from hand-labelled phonetic corpora," *International Journal of Speech Communication*, 29, 1999, pp. 209-224.

Singh, R., B. Raj, and R. Stern, "Automatic generation of subword units for speech recognition systems," *IEEE Transactions on Speech and Audio Processing*, 10, 2002, pp. 89-99.

Soltau, H., F. Metze, C. Fuegen, and A. Waibel, "A One-pass decoder based on polymorphic linguistic context assignment," In *Proceedings of Automctic Speech Recognition and Understanding Workshop*, 2001, Trento, Italy.

Strik, H., and C. Cucchiarini, "Modeling Pronunciation Variation for ASR: Overview and Comparison of Method," *International Journal of Speech Communication*, 29, 1999, pp. 225-246.

Strik H., J.M. Kessens, and M. Wester, "Modeling Pronunciation Variation for Automatic Speech Recognition,", In *Proceedings of the European Speech Communication Association (ESCA) workshop*, 1998, Rolduc, Kerkrade, pp. 137-144.

Torre, D., L. Villarrubia, L. Hernandez, and J.M. Elvira, "Automatic Alternative Transcription Generation and Vocabulary Selection for Flexible Word Recognizers," In *Proceedings of the International Conference on Acoustic*s, *Speech, and Signal Processing*, 1997, Munich, pp. 1463-1466.

Wester, M., and E. Fosler-Lussier, "A comparison of data-derived and knowlegde-based modeling of pronunciation variation," In *Proceedings of International Conference on Spoken Language Processing*, 2000, Beijing, China, pp. 270-273.

Wester, M., "Pronunciation Modeling for ASR knowledge-based, Data-driven Methods," *International Journal of Computer Speech and Language*, 88, 2003, pp. 69-85.

Wester, M., J.M. Kessens, and H. Strik, "Pronunciation Variation in ASR: Which Variation to model?" In *Proceedings of the International Conference on Spoken Language Processing*, 2000, Beijing, China, pp. 488-491.

Yang, Q., and J.-P. Martens, "Data driven lexical modeling of pronunciation variation in ASR," In *Proceedings of the International Conference on Spoken Language Processing*, 2000, Beijing, China, pp. 417-420.

Zeppenfeld, T., M. Finke, K. Ries, M. Westphal, and A. Waibel, "Recognition of conversational speech using the JANUS speech engine," In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 1997, Munich, pp. 1815-1818.