Computational Linguistics and Chinese Language Processing
Vol. 13, No. 2, June 2008, pp. 233-254
© The Association for Computational Linguistics and Chinese Language Processing

Data Driven Approaches to Phonetic Transcription with Integration of Automatic Speech Recognition and Grapheme-to-Phoneme for Spoken Buddhist Sutra

Min-Siong Liang*, Ren-Yuan Lyu+, and Yuang-Chin Chiang#

Abstract

We propose a new approach for performing phonetic transcription of text that utilizes automatic speech recognition (ASR) to help traditional grapheme-to-phoneme (G2P) techniques. This approach was applied to transcribe Chinese text into Taiwanese phonetic symbols. By augmenting the text with speech and using automatic speech recognition with a sausage searching net constructed from multiple pronunciations of text, we are able to reduce the error rate of phonetic transcription. Using a pronunciation lexicon with multiple pronunciations for each item, a transcription error rate of 12.74% was achieved. Further improvement can be achieved by adapting the pronunciation lexicon with pronunciation variation (PV) rules derived manually from corrected transcription in a speech corpus. The PV rules can be categorized into two kinds: knowledge-based and data-driven rules. By incorporating the PV rules, an error rate of 10.56% could be achieved. Although this technique was developed for Taiwanese speech, it could easily be adapted to other Chinese spoken languages or dialects.

Keywords: Automatic Phonetic Transcription, Phone Recognition, Grapheme-to-Phoneme (G2P), Pronunciation Variation, Chinese Text, Taiwanese (Min-Nan), Dialect, Buddhist Sutra.

* Dept. of Electrical Engineering, Chang Gung University, 259 Wen-Hwa 1st Rd., Kwei-Shan Tao-Yuan, Taiwan
E-mail: minsiong@gmail.com
+ Dept. of Computer Science and Information Engineering, Chang Gung University, 259 Wen-Hwa 1st Rd., Kwei-Shan Tao-Yuan, Taiwan
E-mail: renyuan.lyu@gmail.com
Institute of Statistics, National Tsing Hua University, Hsinchu, 101, Section 2 Kuang Fu Rd., 30013, Taiwan
E-mail: chiang@stat.nthu.edu.tw

[Received March 11, 2008; Revised August 4, 2008; Accepted August 11, 2008]

1. Introduction

Automatic phonetic transcription is gaining popularity in the speech processing field, especially in speech recognition, text-to-speech, and speech database construction [Haeb-Umbach *et al*. 1995; Wu *et al*. 1999; Lamel *et al*. 2002; Evermann *et al*. 2004; Nanjo *et al*. 2004; Nouza *et al*. 2004; Sarada *et al*. 2004; Siohan *et al*. 2004; Soltau *et al*. 2005; Kim *et al*. 2005]. It is traditionally performed using two different approaches: an acoustic feature input method and a text input method. The former is the speech recognition task, or more specifically, the phoneme recognition task. The latter is the grapheme-to-phoneme (G2P) task. Both tasks, including phoneme recognition and G2P remain unsolved technology problems. The state-of-the-art speaker-independent (SI) phone recognition accuracy in a large vocabulary task is currently less than 80%, far from human expectations. Although the accuracy of G2P tasks seems much better, it relies on a "perfect" pronunciation lexicon and cannot effectively deal with pronunciation variation issues.

This problem becomes non-trivial when the target text is the Chinese text (漢字). The Chinese writing system is widely used in China and in East/South Asian areas including Taiwan, Singapore, and Hong Kong. Although the same Chinese character is used in different areas, the pronunciation may be very different. Therefore, they are mutually unintelligible and considered different languages rather than dialects by most linguists.

In this paper, we chose Buddhist Sutra (written collections of Buddhist teachings) as the target text processed in this research. Buddhism is a major religion in Taiwan (23% of the population) [IIP 2003]. The Buddhist Sutra, translated into Chinese text in a terse ancient style (古文) , is commonly read in Taiwanese (Min-nan). Due to a lack of proper education, most people are not capable of correctly pronouncing all of the text. Besides, no qualified pronunciation lexicon exists and very few appropriately computational linguistic research projects have been conducted to support developing a G2P system.

Taiwanese uses Chinese characters as a part of the written form, with its own phonetic system differing greatly from Mandarin. This is in contrast to the case of Mandarin, where the problem of multiple pronunciations (MP) is less severe. A Chinese character in Taiwanese can commonly have a classic literate pronunciation (known as Wen-du-in, or "文讀音" in Chinese) and a colloquial pronunciation (known as Bai-du-in, or "白讀音" in Chinese) [Liang *et al*. 2004a]. In addition to MPs, Taiwanese also have pronunciation variation (PV) due to sub-dialectical accents, such as Tainan and Taipei accents. We use the term MPs to stress the fact that variation may cause more deterioration in phonetic transcription [Cremelie *et al*. 1999; Hain 2005; Raux 2004].

The traditional approach to transcribing Chinese Buddhist Sutra text is human dictation. A master monk or nun reads the text aloud, sentence by sentence. Then, some phonetic experts

transcribe the text manually. The manual transcription process is tedious and prone to errors. An example is given in Table 1 as follows [Chen 2006; Tripitaka *et al*. 2005].

Table 1. An example of Transcription of Chinese Buddhist Sutra text into Taiwanese pronunciation, with English translation. The phonetic symbols used here are IPA followed by a digit representing one of several tone classes of the Taiwanese language.

Chinese text of Buddhist Sutra	地藏菩薩本願經：如是我聞。 一時佛在忉利天，為母說法。
Transcription of Taiwanese Pronunciation	tè tsòŋ p’ò sá^{t} pún gùan kíŋ: zù sī ŋó bùn í^{t} sĭ hú^{t} tsài tɤ̄ lì tíen, uì biɤ̂ súa^{t} hūa^{t}
English translation in meaning	Sutra of Earth Treasure: Thus I heard, once the Buddha was in Dao Li Heaven to expound the Dharma to his mother

Since more transcribed Sutras are planned, we are interested in how G2P and ASR technology can help in this situation. Owing to the fact that human experts capable of phonetically transcribing the Sutra in Taiwanese are difficult to find, the first phonetically transcribed Sutra in Taiwanese did not appear until 2004 [Sik 2004a, 2004b]. As shown in Figure 1, our task is to discover which of them is actually pronounced when the Sutra text is segmented into a series of sentences and recorded by a senior master nun. Then, the output of transcription is formed in ForPA or Tongyong Pinin [Lyu *et al*. 2004]. These two phonetic symbol systems are well-designed in ASCII code and suitable for any learners with common understanding of the English phonetic system. This architecture is much easier for a person to use to record his/her reading of the text than acquiring a transcribing expert. For marginalized languages with serious MPs and PV problems, this technique is very useful.

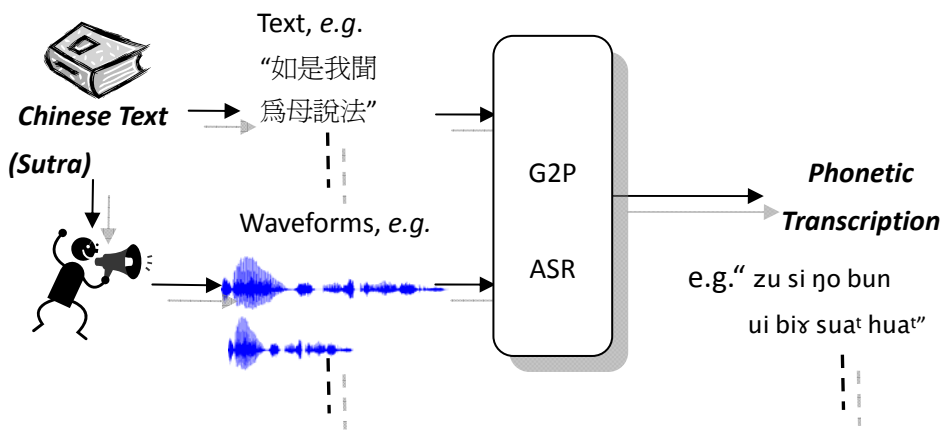

Figure 1. The process of transcribing Chinese text into Taiwanese pronunciation using the ASR technique.

In this paper, we report two experiments using speech and text data, called the Taiwanese Buddhist Sutra (TBS) corpus [Sik 2004b]. The phonetic transcription framework is described in Section 2. Given a speech corpus with phonetic transcription for training, Section 3 reports the speech recognition results with and without the corresponding text for its phonetic transcription. Section 4 discusses the second experiment involving speech recognition with the corresponding text under various pronunciation variation conditions in the training corpus. Section 5 presents our conclusions.

2. The Phonetic Transcription Augmented by Speech Recognition Technique

Figure 2 is the framework of phonetic transcription using the speech recognition technique. While the input is a speech waveform and Chinese Sutra text, the output is a phonetic transcription corresponding to the input Chinese text. The entire framework can be divided into two major parts, *i.e.* an acoustic part and a linguistic part.

Based on flow chart in Figure 2, we define the following notations: $\underline{s}$ is the syllable sequence, while $\underline{c}$ and $\underline{o}$ are the input character and augmented acoustic sequences. The phonetic transcription target is to find the most probable syllable sequence $\underline{s}^*$ given $\underline{o}$ and $\underline{c}$. The formula is:

$$\underline{s}^* = \arg\max_{\forall s \in \underline{S}} P(\underline{s} \mid \underline{o}, \underline{c}) \tag{1}$$

Where $\underline{c} \in \underline{C} = \{\underline{c} \mid \underline{c} = c_1^M = c_1 ... c_M, c_i \in C\}$, c_i is an arbitrary Chinese character, C is the set of all Chinese characters, and the number of elements in C is $n(C)$≈13000. $\underline{s} \in \underline{S} = \{\underline{s} \mid \underline{s} = s_1^N = s_1 ... s_N, s_i \in S\}$, s_i is an arbitrary Taiwanese syllables, S is the set of all Taiwanese syllables, and the number of elements in S is $n(S)$≈1000. Using the Bayes theorem:

$$\underline{s}^* = \arg\max_{\forall \underline{s} \in \underline{S}} \frac{P(\underline{s} \mid \underline{c}) P(\underline{o} \mid \underline{s}, \underline{c})}{P(\underline{o} \mid \underline{c})} \tag{2}$$

The acoustic sequence $\underline{o}$ is assumed dependent only on the syllable sequence $\underline{s}$. Equation 2 could be simplified as:

$$\underline{s}^* = \arg\max_{\forall \underline{s} \in \underline{S}} P(\underline{s} \mid \underline{c}) P(\underline{o} \mid \underline{s}) \tag{3}$$

The first term, $P(\underline{s} \mid \underline{c})$, of Equation 3 is independent of $\underline{o}$ and plays the major role in the linguistic part of the recognition scheme. The second term, $P(\underline{o} \mid \underline{s})$, is the probability of observation given the syllable sequence, which plays the major role in the acoustic part.

For the acoustic part, which is the probability of observing an acoustic sequence $\underline{o}$, given a phonetic syllable sequence $\underline{s}$, it is well known that the Hidden Markov Model (HMM) can be used to model it. We can choose a speaker independent HMM model (SI-HMM) with

speaker adaptation techniques.

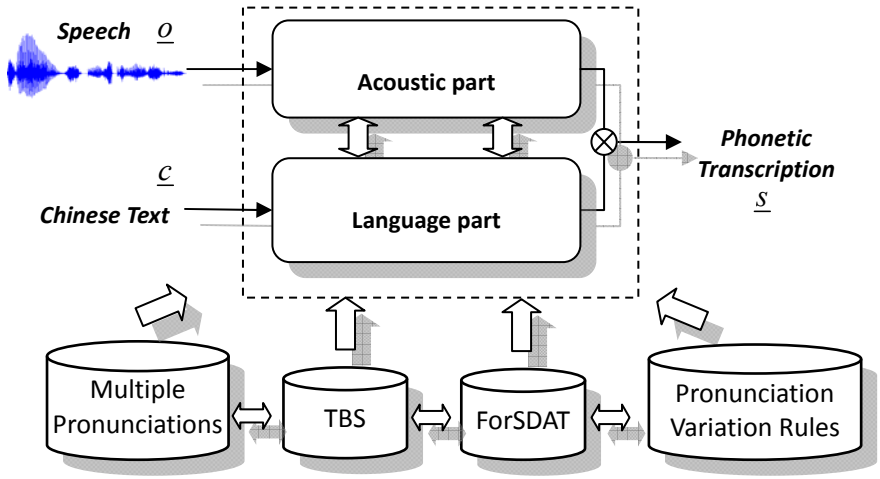

Figure 2. The flow chart of the phonetic transcription of Taiwanese Buddhist Sutra (TBS) incorporating pronunciation variation rules.

The linguistic part, which is the probability of observing a syllable sequence $\underline{s}$, given a character sequence $\underline{c}$, could be modeled as a traditional grapheme-to-phoneme problem. In such a problem, a "well-coverage" phonetic lexicon, which covers as many as possible correct pronunciations for each phoneme, is quite useful. The problem of multiple pronunciations could be solved using a specially designed searching net, such as the sausage net, which was named for its shape being similar to a sausage. All the searching nets, including the sausage net, were constructed according to a multiple pronunciation lexicon and described in the next section. Even the best pronunciation lexicon would miss the true pronunciation for a certain Chinese character. This is severe, especially for a minority language without many linguistic resources, like Taiwanese. To address this issue, the pronunciation variation rules would be incorporated in a sausage net to improve the accuracy of transcription.

3. Solutions to Multiple Pronunciation Problem

The first experiment is performed on the Sutra phonetic transcription using the sausage recognition network without considering the pronunciation variation problem. For a syllabic language such as Taiwanese or Mandarin, we can construct a concatenated net of all syllables. Based on Equation 3, we define: $\underline{s} = s_1, s_2, ..., s_N$ as the syllable sequence. As our goal is to find the real pronunciation, it would not be crucial to know the relationship between Chinese characters and syllables. Therefore, assume that the underlined character sequence $\underline{c}$ is known and independent of $\underline{s}$, and all syllables are independent of each other. Following Equation 3, we have:

$$\begin{aligned}\underline{s}^{*} &= \arg\max_{\forall \underline{s}\in\underline{S}} P(\underline{s})P(\underline{o}\mid\underline{s}) \\ &= \arg\max_{\forall \underline{s}\in\underline{S}} P(s_1)P(s_2)...P(s_N)P(\underline{o}\mid s_1,s_2,...,s_N) \\ &= \arg\max_{\forall \underline{s}\in\underline{S}} P(\underline{o}\mid s_1,s_2,...,s_N)\prod_{i=1}^{N}P(s_i)\end{aligned} \tag{4}$$

To make the case simple and straightforward, we could assume that $P(s_i)$ is a uniform distribution, then Equation 4 can be simplified as follows:

$$\underline{s}^{*} = \arg\max_{\forall \underline{s}\in\underline{S}}(\frac{1}{n(S)})^{N} P(\underline{o}\mid s_1,s_2,...,s_N) = \arg\max_{\forall \underline{s}\in\underline{S}} P(\underline{o}\mid s_1,s_2,...,s_N) \tag{5}$$

where S is the set of all possible syllables and the *n(S)* is the number of elements in *S*.

Such a concatenated net is called a total-syllable net. It is a compact representation of the searching space $\underline{S}$, which is a set of all possible syllable sequences. The transcription performance in this way is dependent only on the acoustic part. Therefore, the experimental results conducted using a total-syllable net is referred to as the performance of the acoustic part.

Second, it is also possible to perform the phonetic transcription using only text input without any speech/acoustic clues. This is the linguistic part in the recognition scheme shown in Fig. 2. In this case, Equation 3 can be simplified as:

$$\underline{s}^{*} = \arg\max_{\forall \underline{s}\in\underline{S}} P_{\underline{S}|\underline{C}}(\underline{s}\mid\underline{c}) = \mathrm{P}_{S_1...S_N|\mathrm{C}_1...\mathrm{C}_N}(s_1...s_i...s_N \mid \mathrm{c}_1...\mathrm{c}_i...\mathrm{c}_N) \tag{6}$$

As only a small scale database is available, we assume that s_i is dependent on c_i and s_{i-1}, or even only on c_i. Equation 6 can then be simplified as:

$$\underline{s}^{*} = \arg\max_{\forall s_i\in S_i} \prod_{i=1}^{N} P_{S_i|C_i}(s_i \mid c_i) \tag{7}$$

and

$$\underline{s}^{*} = \arg\max_{\forall s_i\in S_i} \prod_{i=1}^{N} P_{S_i|C_i}(s_i \mid s_{i-1}, c_i) \tag{8}$$

The results from the experiments conducted using Equations. 7 and 8 depend only on the textual input instead of the acoustic input, and are referred to as the language part performance. Therefore, discussion about Equations 5 and 7 require traditional automatic speech recognition and grapheme-to-phoneme approaches for dealing with the phonetic transcription tasks.

What is proposed in this paper is an approach to integrate both. Given a Chinese character sequence, based on the multiple pronunciations of each Chinese character, a much smaller recognition net can be constructed. Thus, by integrating Equations 5 and 7, we have:

$$\underline{s}^* = \arg\max_{\forall s_i \in S_i} P(\underline{o} \mid s_1, s_2, ..., s_N) \prod_{i=1}^{N} P(s_i) P_{S_i|C_i}(s_i \mid c_i) \tag{9}$$

Taking an example of a typical text sentence "爲母說法", which is shown in Figure 3, we will call such a net (with multiple pronunciations) a sausage net. Higher recognition accuracy can be expected due to the smaller perplexity in the recognition net. Our task is to construct "good" sausage nets to help the acoustic part do the job. In the following, we will discuss how to use the lexicons, the recognition networks to implement the proposed framework and show some experiment results.

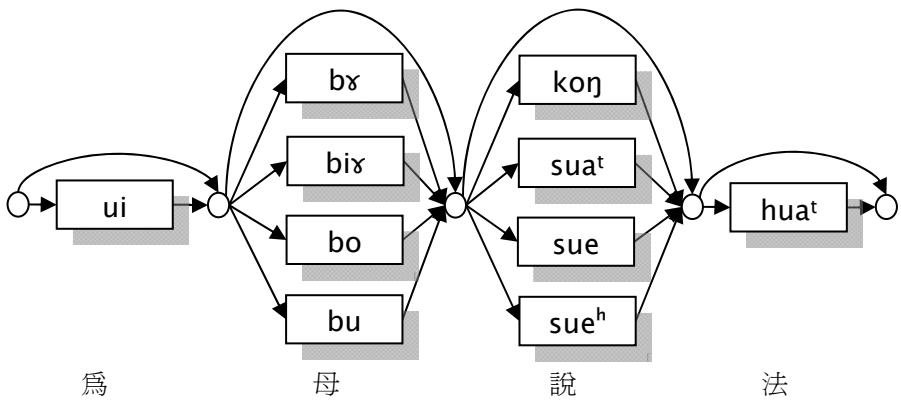

Figure 3. The sausage searching net. The net is constructed from the multiple pronunciations of each Chinese character from the Formosa Lexicon. The corresponding Chinese characters with multiple pronunciations are also shown.

3.1 Speech Database

In this paper, we use as the speech database, Formosa Speech Database (ForSDAT), which was collected over the past several years [Lyu *et al*. 2004]. The SI-HMM model can be trained from the ForSDAT-01, which contains 200 speakers and 23 hours of speech. All speech data were recorded in 16K, 16bit PCM format. The statistical information of ForsDAT-01 was summarized as in Table 2.

In addition, the partial ForSDAT-02 speech corpus was used to derive the rule set of pronunciation variations, which contains 131 speakers and 7.2 hours of speech. The statistical information of partial ForsDAT-02 was summarized as in Table 2 and the detail is discussed in Section 4.

The distribution of another speech database, TBS, is listed in Table 3, where there are 1,619 utterances in this speech data set with a total length of about 230 minutes [Sik 2004b]. 502 utterances, which include 5909 syllables, are randomly chosen and reserved for testing

while as another 31 utterances are used for acoustic model development.

Table 2. The statistics of ForSDAT-01 speech corpus and partially manually validated ForSDAT-02 speech corpus.

	ForSDAT-01	Partial ForSDAT-02
Utterance	92158	19731
Number of People	100 (male: 50, female: 50)	131 (male: 72, female: 59)
Number of Syllable	179730	45865
Number of Distinct Triphones	1356	1194
Number of Total Triphones	555731	104894
Time (hr)	22.43	7.2

Table 3. TBS (Taiwanese Buddhist Sutra) speech corpus.

Buddhist Corpus Category	Utterance	# of Syllable	Time (min)
Adaptation	31	359	2.62
Test	502	5909	43.23
Other	1086	12147	179.88
Total	**1619**	**18415**	**225.73**

3.2 The Pronunciation Lexica

There is one pronunciation lexicon available to us, **Formosa Lexicon**, which provides multiple pronunciations in Taiwanese for all Chinese characters. The lexicon contains about 123,000 words in Chinese/Taiwanese text with Mandarin/Taiwanese pronunciations. It is a combination of two lexica: Formosa Mandarin-Taiwanese Bi-lingual lexicon and Gang's Taiwanese lexicon [Liang *et al*. 2004a; Liang *et al*. 2004b; Lyu *et al*. 2004]. The former is derived from a Mandarin lexicon; thus, many commonly used Taiwanese terms are missing due to the fundamental difference between these two languages. The latter contains more widely used Taiwanese expressions from samples of radio talk shows. Some examples of the lexicon are shown in Table 4, containing 65,007 entries using the Wen-du-in pronunciation and 58,431 entries using the Bai-du-in pronunciation. There are a total of 123,438 pronunciation entries. For all 65,007 Wen-du-in pronunciation entries, there are 6,890 entries for one-syllable words, 39,840 entries for two-syllable words, and so on. The lexicon as described above is a general-purpose lexicon. It could be used for a wide range of applications and tends to have a higher number of multiple pronunciations.

We used two kinds of searching nets in these experiments, according to the lexicon. The first is the Total-syllable net. It is simply a concatenated net of all Taiwanese syllables existing in the Taiwanese Buddhist Sutra (TBS), where the total number of syllables is 467, denoted as the Total-Syl-Net. The other searching nets are the sausage nets generated from

each of the pronunciation lexica. The nets were constructed by filling in each node of the net with the corresponding multiple pronunciations of each Chinese character from the pronunciation lexicon. One example is shown in Figure 3. The nets are denoted the General-Sau-Net for the general-purpose Formosa Lexicon.

Table 4. The partial example of all possible pronunciations per Chinese Character from Formosa bi-lingual Lexicons, including classic literature pronunciation (Wen-du-in) or daily life pronunciation (Bai-du-in).

	Pronunciation 1	**Pronunciation 2**	**Pronunciation 3**	**……**
日(sun)	gí^{p}	lí^{p}	zí^{p}	……
火(fire)	xê	xùe		
加(add)	ká	ké	kúe	……
叩(knock)	k'áu	kìɤ	k'ō^{k}	……
卵(egg)	lī̍	lûan	nī̍	……
坐(sit)	tsé	tsɤ̄	tsūe	……

However, a lexicon is inevitably incomplete, and we could be confronted with the missing character problem and the missing pronunciation problem. The missing character problem is when a character used in the Sutra does not appear in the lexicon. One reason is because many of the Chinese characters used in ancient times are no longer used in modern times. Thus, even the Unicode Standard, which contains more than thirty thousand Chinese characters, does not contain them. The Formosa Lexicon has much fewer distinct characters, and the missing character problem is inevitable. When a missing character is encountered, we use all possible syllables as its multiple pronunciations. One example is illustrated in Figure 4, where the sausage searching net is constructed for the Chinese character string "$C_0C_1C_2$". It is assumed that the character C_0 is a missing character. In such a case, all possible syllables, denoted as $S_{00}, S_{01}, ...S_{0N}$, are used as possible pronunciations of C_0.

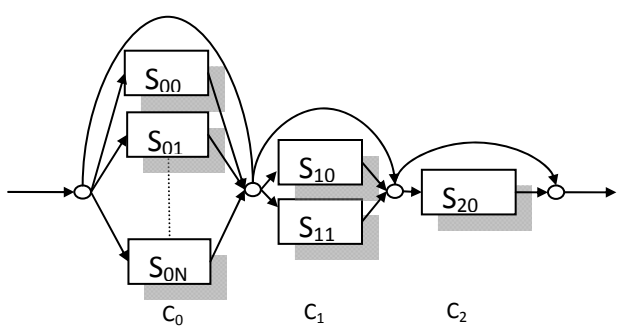

Figure 4. The sausage searching net with missing character C_0, where all syllables, S_{00}, S_{01},..., S_{0N}, are used as its possible pronunciations.

4. Incorporating Pronunciation Variation Rules

As insufficient coverage of pronunciation nodes in the searching net will severely degrade the recognition performance, some approaches to extend the pronunciation coverage will be considered to help the overall performance. Since global lexicon modification by experts would take considerable effort and not necessarily benefit, we adopted alternative rule-based methods. By rule-based pronunciation variations, we mean that phonetic units will be changed by speakers according to some underlying rules. Usually, a rule could be notated as the form "*B*→*S*" for canonical pronunciation *B* (base-form) being substituted with the actual pronunciation *S* (surface-form) [Saraclar *et al*. 2004]. Briefly speaking, some rule-derived variant pronunciations are added directly into the searching net to enhance the poor pronunciation coverage of an imperfect pronunciation lexicon.

An example is shown in Figure 5, where the number of pronunciations for the Chinese character "母" was increased from 4 to 5 by incorporating some specific pronunciation rules as "/ɤ/ → /o/". It could be shown that, as long as the pronunciation rules could be well designed, the phonetic transcription performance would be effectively improved.

Generally speaking, the pronunciation-variation (PV) rules can be categorized into two kinds: knowledge-based and data-driven rules. The knowledge-based rules were derived from the knowledge established by phoneticians. On the other hand, the data-driven PV rules rely on the availability of transcribed speech corpora.

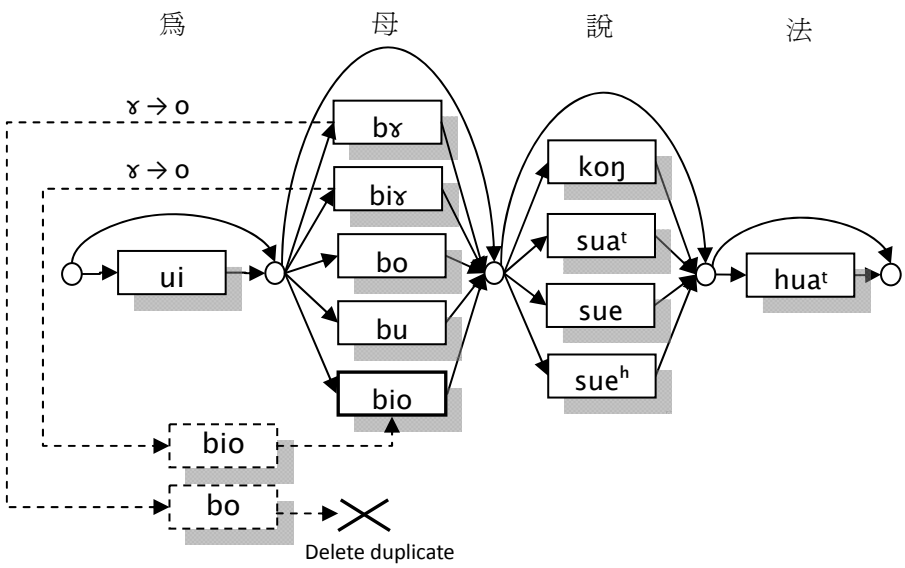

Figure 5. An example of the extended sausage searching net. The net is constructed from the multiple pronunciations in lexicon and expanded using pronunciation-variation rules for each Chinese character according to the rule "/ɤ/ → /o/".

4.1 The Knowledge-Based Variation Rules

Considering the trade-off between the number of elements and the degree of detail or perplexity, the triphone was used as the acoustic unit, thus, the transcription unit in this paper. The form *LBR* → *LSR* represents the pronunciation variation rules, where *B* and *S* represent the base form and surface form of a central phone, and *L*, *R* are the left and right contexts respectively. The number of triphone units in Taiwanese is about 1200.

As with the other members of the Chinese language family, there are about three types of pronunciation variations in Taiwanese. These could be summarized as follows and are shown in Table 5:

1. Variation between Bai-du-in and Wen-du-in: The variations may vary due to using classic literate pronunciation (known as Wen-du-in) or a colloquial pronunciation (known as Bai-du-in). This point has been discussed previously in Section 1. For example, the Chinese character "生" (to give birth) might be pronounced as /siŋ/ in Wen-du-in and /sẽ/ or / sĩ / in Bai-du-in. All of these are acceptable.
2. Variation between sub-dialectal regions: Some variations were referred to as dialectal differences; For instance, the initials /z/ is substituted with /l/ or /g/ depending on the sub-dialect of Taiwanese. Such a rule was denoted as "z→l/g".
3. Variation due to personal pronunciation errors: Some kinds of variations are considered personal pronunciation errors. Owing to the lack of some phonemes in the mainstream language (such as Mandarin in Taiwan), some pronunciation may disappear in younger generations. One of these phonemes is /g/, where the phenomenon is denoted as "g → {}".

Table 5 can be considered as a knowledge source to select the pronunciation variation rules. The knowledge-based PV rules, which were derived by more than one linguist, were sometimes contradictory with each other. This made them difficult to choose between at times in implementation. Such a difficulty leads to a need for another approach. One of them is a data-driven approach from large-scale real data. Since a little manually transcribed speech data was available, we could use statistical computational measures to extract the PV rules from real data. This issue will be discussed in the following section.

Table 5. The three types of pronunciation variations in Taiwanese.

Variation types	Examples
Bai-du-in / Wen-du-in	siŋ → sẽ/sĩ
Dialectal difference	z → l/g
	ɤ ↔ɔ → o
	n →l
	b → m
	ĩũ → ĩõ
Personal pronunciation error	g →{}
	b →{}
	h →{}

4.2 The Sata-Driven Variation Rules

The same simple way to adopt the methodology of pronunciation variation is to expand the pronunciation lexicon using variation rules of the form *LBR* → *LSR*. Similar work for such an approach was shown in Mandarin [Tsai *et al*. 2002]. To derive such rules, a speech corpus with both canonical pronunciation and actual pronunciation is necessary. We choose a subset of ForSDAT, called ForSDAT-02, to derive PV rules, and the statistical information is summarized as in Table 2.

ForSDAT-02 is a speech database with rich bi-phone coverage. This database was recorded by prompting speakers with a script. Although the script in Taiwanese text was shown with phonetic transcription, we did observe variations in the recorded speech. A small portion of the speech data was then manually checked, and the phonetic transcription of the script was corrected according to actual speech. Some examples of the original transcription (the base-form) and the manually corrected transcription (the surface-form) are shown in Table 6, which is called the sentence-level confusion table.

Table 6. Sentence-level confusion table. The output is manually corrected transcription (the surface-form), and the input is the original transcription (the base-form).

Original transcription (base-form)	Manually corrected transcription (surface-form)
è bɤ̀ k'î ă	è bŏ k'î ă
gám k'ài bàn ts'én	gán k'ài bàn ts'én
sĭŋ ùa^{h} hù^{h} hŭa^{h}	sĭn ùa^{h} hù^{h} hŭa^{h}
……	……

From the sentence-level confusion table, it is quite a straightforward process to construct other confusion tables in syllable level and triphone level. These two tables are shown in Table 7 and Table 8 as follows.

Table 7. Syllable-level confusion table, where z_{ij} represents the number of variation from syllable x_i (base-form) to triphone y_j (surface-form), T is the number of surface-form and base-form

	bɤ	bo	……	y_j	……	siŋ	sin
bɤ	237	30	……	z_{1j}	……	0	0
bo	0	64	……	z_{2j}	……	0	0
.....	...	...	……	...	……		
x_i	z_{i1}	z_{i2}	……	z_{ij}	……	$z_{i,T-1}$	z_{iT}
.....	...	...	……	...	……		
siŋ	0	0	……	$z_{T-1,j}$	……	163	12
sin	0	0	……	z_{Tj}	……	2	105

The triphone-level confusion table is used as a direct knowledge source to derive the PV rules, where each cell in the table was looked upon as a rule. The number of rules shown in Table 8 is P^2, where P is the number of triphones (about 1200 in the target language). The number of rule set selections is 2^{P^2}, which is an enormous number which is impossible to be processed in modern computers. To make the problem more solvable, some specially designed algorithms should be developed that are able to specifically find a useful route in the huge rule set selection space within a reasonable time.

Table 8. Triphone-level confusion table, where n_{ij} represents the number of variation from triphone b_i to triphone s_j, P is the number of surface-form and base-form, $N_i = \sum_j n_{ij}$, $M_j = \sum_i n_{ij}$ ***and*** $N = \sum_i \sum_j n_{ij}$

	bh-er	i-ng	*i-n*	……	s_j	……	*bh-o*	a-n	*a-m*	
bh-er	237	0	0	……	n_{1j}	……	30	0	0	267
i-ng	0	1273	84	……	n_{2j}	……	0	0	0	1373
…	…	…	…		…		…	…	…	
b_i	n_{i1}	n_{i2}	n_{i3}	……	n_{ij}	……	$n_{i,p-2}$	$n_{i,p-1}$	n_{iP}	N_i
…				……		……				
a-m	0	0	0	……	n_{Pj}	……	0	35	834	
	241	1315	1102		M_j		107	1873	870	N

First of all, some criteria should be adopted to choose the most significant rule sets. Three kinds of statistical measures were used in this paper. They are (1) Joint probability [Raux 2004], (2) Conditional probability, and (3) Mutual information-like of the base form pronunciation and the surface form pronunciation. The mathematical definitions of the above three measures are as follows:

1. Joint probability of the base form pronunciation b_i, and the surface form pronunciation s_j,
   $p(b_i, s_j) = n_{ij} / N$
2. Conditional probability of the surface form pronunciation s_j, conditioned on the base form pronunciation b_i,
   $p(s_j \mid b_i) = n_{ij} / N_i$
3. Mutual information of the base form pronunciation b_i, and the surface form pronunciation s_j,
   $$I_{ij} = p(b_i, s_j) log \frac{p(b_i, s_j)}{p(b_i) p(s_j)} = \frac{n_{ij}}{N} log(N \frac{n_{ij}}{\sum_i n_{ij} \cdot \sum_j n_{ij}})$$

In all of the above equations, n_{ij} is the number of (base-form) triphone b_i substitutions by the surface-form triphone s_j that appear in a corpus, and

$N = \sum_i \sum_j n_{ij}$,

$N_i = \sum_j n_{ij}$,

$p(b_i, s_j)$ represents the joint probability of (b_i,s_j),

$p(b_i)$ and $p(s_j)$ equal the marginal probability of b_i and s_j, *respectively.*

Note that each pair *(i,j)*,$i \neq j$, corresponds to a substitution rule, and we select those pairs *(i,j)* with higher scores of $p(b_i, s_j)$, $p(b_i, s_j)$ and I_{ij} to be the variation rules to extend the sausage net pronunciation.

In Table 9, the rules were sorted by rank based on joint probability, conditional-probability, and mutual-information. There are variants among the three lists. One rule which is much more important in some method may be trivial in the other method.

Table 9. Data-driven rules: The top 10 substitution errors were listed from the partially validated ForSDAT-02 corpus for Joint-Probability-Based, Conditional-Probability-Based and Mutual-information-Based method

Rank based on Joint Probability	Rank based on Conditional Probability	Rank based on Mutual-Information
i-ŋ → i-n	x-ã^{h} → x-a^{h}	i-ŋ → i-n
a-m → a-n	**n-ɤ → n-õ**	**b-ɤ → b-o**
b-ɤ → b-o	ŋ-ĩ → ŋ-ĕ	ĩ-õ → ĩ-ũ
i-m → i-n	l-o^{h} → l-o	l-i^{k} → l-i^{t}
ĩ-õ → ĩ-ũ	k'-i^{h} → k'-i^{k}	**k'-ɤ → k'-o**
a-n → a-m	ts-a^{k} → ts-a^{t}	ĩ-ŋ → ĩ-n
a-ŋ → a-m	**g-ɤ → g-o**	**p-ɤ → p-o**
i-a-ŋ → i-o-ŋ	p'-i^{k} → p'-i^{t}	i-m → i-n
i-m → i-ŋ	ĩ-õ → ĩ-ũ	**t-ɤ → t-o**
i-n → i-m	g-i^{k} → g-i	b-i^{t} → b-i^{h}

5. The Experiment Results and Discussion

In training or estimating SI-HMM models for the acoustic part, we use continuous Gaussian-mixture HMM models with feature vectors of 12-dimensional MFCC with 1-dimensional energy, plus the first, second, and third derivatives computed using a 20-ms frame width and 10-ms frame shift. Context-dependent intra-syllabic tri-phone models were built using a decision-tree state tying procedure. As the testing data is speaker dependent, adaptation with some manually transcribed data must be useful in automatic phonetic

transcription. Maximum Likelihood Linear Regression (MLLR) is then used to adapt speaker independent models using 31-utterance adaptation speech data. Most of the training and recognition are carried out by using the HTK tools [Young *et al*. 2008].

With the two searching nets (Total-Syl-Net, General-Sau-Net) and acoustic models (SI with adaptation), the recognition results measured as the syllable error rate (SER) are shown in Figure 6. In addition, we also show the result of only language, called grapheme-to-phoneme (G2P), with unigram.

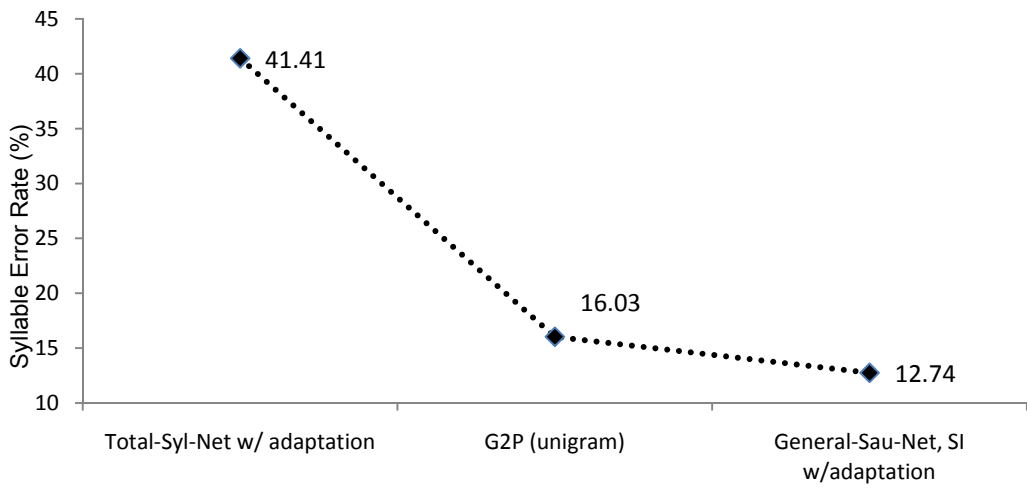

Figure 6. Syllable error rate (SER) under uni-gram, and General-Sau-Net, SI w/ adaptation. See text in subsection 3.2 for notations.

Through observation of the experimental results, we can see that neither G2P with unigram nor Total-Syl-Net with adaptation model can reach acceptable performance. Therefore, it is necessary to integrate the linguistic and acoustic parts. General-Sau-Net could surpass Total-Syl-Net. For example, under the same speaker adaptation models, the result was 12.74% better with General-Sau-Net than other results in Figure 6. Thus, if the speaker independent model could be adapted using some phonetically transcribed speech data, the adapted speaker independent model under General-Sau-Net would be suitable for the phonetic annotation task. In multiple pronunciation problems, by our speech data observation, we could see that some errors result from pronunciation variations. Therefore, we hypothesized that the performance would get better by adaptation of the Formosa Lexicon Sausage net, *i.e*. adaptation of the Formosa lexicon, described in the Section 4.

We adapted the Formosa (general-purpose) pronunciation lexicon according to different pronunciation variation rule sets. The speech recognition task with a sausage searching net and speaker adapted acoustic models was then conducted, as described in Section 4, wherein, the

SER achieved before the application of the pronunciation variation rules was **12.74%**, as shown in Figure 6. This would be looked upon as the performance of the baseline setup in this section.

In Figure 7, the transcription performance was measured in terms of syllable error rate vs. the number of ranked PV rules sorted according to different measures, including mutual-information (MI), joint-probability (JP), and conditional-probability (CP) as well as the baseline setup. We could observe that it is truly helpful to decrease the SER by increasing the searching net coverage via the PV rules. The evidence is that the lowest error rate (**10.56%**) was achieved by utilizing the first 52 variation rules, which were selected by the Mutual-Information (MI) measure. Similar improvement would also be observed in the best SER (**11.81%** and **11%**) achieved using the Joint-Probability (JP) and Conditional-Probability (CP) measures when the complexities of the JP and CP measure are 2.68 and 2.55, respectively.

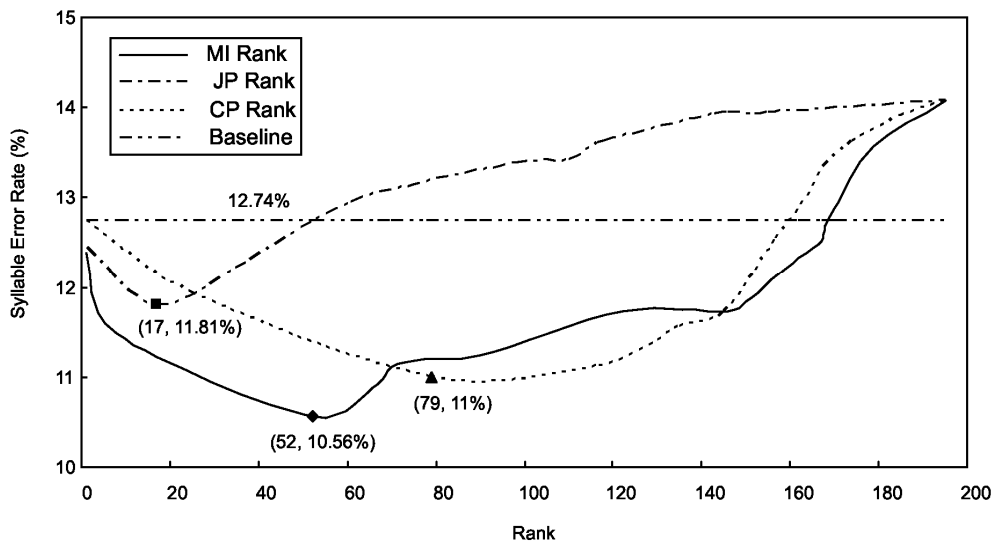

Figure 7. The recognition result (Syllable error rate) v.s. the number of ranked rules sorted according to different measures, including mutual-information (MI), joint-probability (JP), and conditional-probability (CP) as well as the Baseline criterion.

It is interesting to point out that, in Figure 7, choosing different statistical measures was found to influence the achievable lowest SER and also the speed of decrease in SER. In these experiments, we found that MI was the best in terms of the rate of decrease in SER or the achievable lowest SER. Although the JP-based measure could make the error rate converge more quickly than the CP-based measure, the performance also degraded quickly. This is because the CP-based measure score was normalized by the base-form count in contrast to the

JP-based measure. However, the insignificant and harmless PV-rules might get the higher conditional probability sometimes due to few base-form observations. The PV-rules for the CP-based measure might not increase the perplexity but still lead to the slowest convergence among the three measures. In the MI-based measure, the formula could avoid slow convergence using the Joint-Probability as a weight when the base-form had few variations. Observing the confusion table, the surface-form would have lower correlation with these base-forms if many base-forms would transform into the same surface-form. So, we proposed that the mutual information between the base and surface-forms should be used to calculate the base and surface-form correlation using the normalization of their count. Consequently, the error rate of the MI rank converges most quickly and the performance of the MI measure in error reduction was also better than JP and CP measures, respectively.

Another interesting point was that the SER will possibly increase if too many PV rules are applied. For example, the lowest SER is achieved by applying 52 rules when MI was adopted as the ranking measure. However, after applying more rules, the SER increased! It even became worse than that in the baseline experiments. This means that some "bad" pronunciation variation rules may lead to a performance reduction. Take the Joint-Probability (JP) measure for example. The optimal performance was achieved when 17 ranked rules were applied, but when the number of rules further increased, the performance degraded. It was similar when MI or CP was used. Therefore, it is important to determine "good" rules and choose them so that the optimal performance could be achieved as soon as possible.

Extending the searching net can enhance the SER performance, but the extension must be limited to a suitable range. This point can be observed from the perplexity of the searching net in Figure 8. Regardless what measures we use, the differences in the perplexity values from the best results among the three measures were always slight. For example, in Figure 8, the perplexity of the best JP measure result was 2.68 when the perplexity of best MI measure result was 2.62. That means too many rules may lead to more real pronunciation coverage, but the performance may improve slightly or even decrease progressively. The perplexity is a good measure to evaluate the searching net in obtaining the best results.

Finally, in Figure 9, the error rate of General-Sau-Net is 12.74%. However, some errors resulted from pronunciation variations caused by a speaker's accent. Therefore, through incorporating variation rules into General-Sau-Net with different statistic measures, the best error rates can be reduced to 11.81%, 11%, and 10.56% with respect to JP-, CP-, and MI-based measures, respectively.

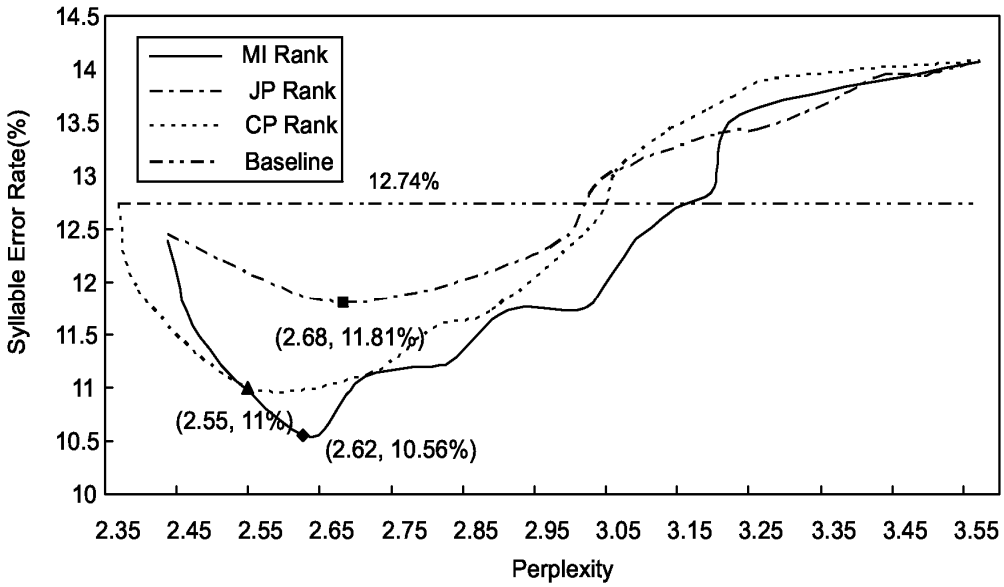

Figure 8. The recognition result (syllable error rate) vs. the perplexity sorted according to different measures, including mutual-information (MI), joint-probability (JP), and conditional-probability (CP) as well as the Baseline criterion

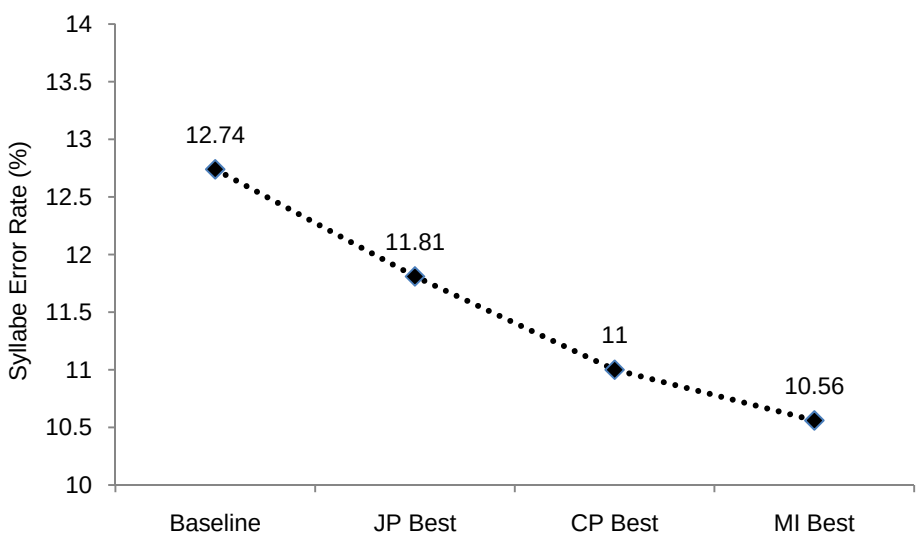

Figure 9. The recognition result (syllable error rate) under four kinds of net according to different measures, including mutual-information (MI), joint-probability (JP), and conditional-probability (CP) as well as the Baseline criterion.

6. Conclusions

We have proposed a new approach to address the phonetic transcription of Chinese text into Taiwanese pronunciation. Considering the fact that there are very few linguistic resources for Taiwanese, we used speech recognition techniques to deal with multiple pronunciation variations, which is a very common phenomenon in Taiwanese but hard to deal with using traditional text-based approaches. A general-purpose lexicon (called the Formosa lexicon), and a speaker-adapted HMM model were used to achieve a syllable error rate, 12.74%. In order to enhance the performance, the trivial adaptation of a general-purpose sausage net with pronunciation variation rules was used instead of global pronunciation lexicon modification.

In pronunciation variation rule (PV-rules) selection, the data-driven variation rules, which were derived using three statistical measures, were used to extend more possible pronunciations. Although the knowledge-based rules were also derived from a knowledge source, the rules were difficult to implement and dependent on the specific language. Thus, we selected data-driven rules with context-dependent triphones as the general solution to the PV problem. In the data-driven measure, the mutual-information-based (MI) rank outperformed the Joint-Probability rank and Conditional-Probability rank. Compared with baseline experiment result error rate of 12.74%, the lowest error rate of the MI-based measures had an error reduction rate of 17.11%, which was the best among the three statistical measures proposed in this paper. The error rate of the MI rank converged most quickly and the best performance of MI-based measure appeared in the first 52 ranks.

The experimental results from data-driven measures could possibly provide the evidence to help choose the corresponding knowledge-based PV rules. Of course, some of the pronunciation variation rules were certainly language-dependent (*i.e*. the phonological and phonetic processes differ between languages) [Kanokphara *et al*. 2003]. However, the major points to be emphasized are that the proposed technique to model pronunciation variation for transcription was rather language-independent.

The recognition of tones was still an unsolved problem in this research. This is another issue for further research. In Taiwanese, there are 7 tone classes, which could be used to distinguish the meanings of words. In addition, the complex tone sandhi would also be accompanied with tone recognition. If more speech and text was gathered, the analysis and statistics of pronunciation information for pronunciation probability would be the next step. We will construct a human interaction system to help more Taiwanese publications be presented. This technology may also be used as a language-learning tool.

Although the proposed technique was developed for Taiwanese speech, it could also be easily adapted for application in other similar “minority” Chinese spoken languages, such as Hakka, Wu, Yue, Xiang, Gan, and Min, or other non-Han family languages which also use

Chinese characters as the written language form.

In summary, the proposed semi-automatic transcription of Chinese text into a Taiwanese pronunciation system reached a **12.74%** error rate in the baseline experiment. Further improvement using pronunciation variation rules produced a **17.11%** error rate reduction.

Reference

Chen C. H., Sutra on the original Vows of Bodhisattva Earth Treasure in English, http://www.yogichen.org/efiles/b041a.html, 2006.

Cover, T. M. and J. A. Thomas, *Elements of Information Theory*, New York: Wiley, 1991.

Cremelie, N., and J.-P. Martens, "In Search of Better Pronunciation Models for Speech Recognition," *Speech Communication*, 29, 1999, pp. 115-136.

Evermann, G., H.Y. Chan, M.J.F. Gales, T. Hain, X. Liu, D. Mrva, L. Wang, and P.C. Woodland, "Development of the 2003 CU-HTK Conversational Telephone Speech Transcription System," *In: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2004, Montreal, Canada, pp. I-249-I-252.

Haeb-Umbach, R., P. Beyerlein, and E. Thelen, "Automatic Transcription of Unknown Words in a Speech Recognition system," *In: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 1995, pp. 840-843.

Hain, T., "Implicit modeling of pronunciation variation in automatic speech recognition," *Speech Communication*, 46, 2005, pp. 171-188.

U.S. Department of State's Bureau of International Information Programs, IIP report, http://usinfo.state.gov/, Dec, 2003.

Kanokphara, S., V. Tesprasit, and R. Thongprasirt, "Pronunciation Variation Speech Recognition without Dictionary Modification on Sparse Database," *In: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2003, Hong Kong, pp. I-764-I-767.

Kim, D. Y., H.Y. Chan, G. Evermann, M.J.F. Gales, D. Mrva, K.C. Sim and P.C. Woodland, "Development of the CU-HTK 2004 Broadcast News Transcription Systems," *In: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2005, Philadelphia, USA, pp. 861-864.

Lamel, L., J.-L. Gauvain, and G. Adda, "Lightly Supervised and Unsupervised Acoustic Model Training," *Computer Speech and Language,* 16, 2002, pp. 115-129.

Liang, M.-S., R.-C. Yang, Y.-C. Chiang, D.-C. Lyu and R.-Y. Lyu, "A Taiwanese Text-to-Speech System with Applications to Language Learning," *In: Proc. Int. Conf. on Advanced Learning Technologies (ICALT)*, 2004, Joensuu, Finland, pp. 91-95.

Liang, M.-S., D.-C. Lyu, Y.-C. Chiang and R.-Y. Lyu, "Construct a Multi-Lingual Speech Corpus in Taiwan with Extracting Phonetically Balanced Articles," *In: Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, 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 & Chinese Language Processing (IJCLCLP)*, 9(2), August 2004, pp. 1-12.

Nanjo, H., and T. Kawahara, "Language Model and Speaking Rate Adaptation for Spontaneous Presentation Speech Recognition," *IEEE Transaction on Speech and Audio Processing*, vol. 12, Jul. 2004, pp. 391-400.

Nouza, J., D. Nejedlova, J. Zdansky and J. Kolorenc, "Very Large Vocabulary Speech Recognition System for Automatic Transcription of Czech Broadcast Programs," *In: Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, 2004, Jeju, Korea.

Raux, A., "Automated Lexical Adaptation and Speaker Clustering based on Pronunciation Habits for Non-Native Speech Recognition," *In: Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, 2004, Jeju Island, Korea.

Saraclar, M., and S. Khudanpur, "Pronunciation change in conversation speech and its implications for automatic speech recognition," *Computer Speech and Language*, 18, 2004, pp. 375-395.

Sarada, G.L., and N. Hemalatha, T. Nagarajan and Hema A. Murthy, "Automatic Transcription of Continuous Speech using Unsupervised and Incremental Training," *In: Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, 2004, Jeju, Korea.

Sik, D.-G., *The Four Basic Sutra in Taiwanese*, DiGuan Temple, HsinChu, Taiwan, 2004.

Sik, D.-G., *Earth Treasure Sutra in Taiwanese*, DiGuan Temple, HsinChu, Taiwan, 2004.

Siohan, O., B. Ramabhadran, and G. Zweig, "Speech Recognition Error Analysis on the English MALACH Corpus," *In: Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, 2004, Jeju Island, Korea.

Soltau, H., B. Kingsbury, L. Mangu, D. Povey, G. Saon and G. Zweig, "The IBM 2004 Conversational Telephony System for Rich Transcription," *In: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2005, Philadelphia, USA, pp. I-205-I-208.

Tripitaka, S. S., Sutra on the original Vows of Bodhisattva Earth Treasure in Chinese. http://book.bfnn.org/article/0016.htm, 2005.

Tsai, M.Y., F.C. Chou, and L.S. Lee, "Improved pronunciation modeling by inverse word frequency and pronunciation entropy," *In: Proc. IEEE Int. Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2002, pp. 53-56.

Wu, J., and V. Gupta, "Application of Simultaneous Decoding Algorithm to Automatic Transcription of Known and Unknown Words," *In: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 1999, Phoenix, USA, pp. 589-592.

Young, S., G. Evermann, M. Gales, T. Hain, D. Kershaw, X. (Andrew) Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. Woodland, *The HTK Book*, 3.4 ed., 2008.