The Multiple Pronunciations in Taiwanese and the Automatic Transcription of Buddhist Sutra with Augmented Read Speech

Yuang-Chin Chiang[+], *Min-Siong Liang*[*], *Hong-Yi Lin*[†], *Ren-Yuan Lyu*[†]

[+]Inst. of Statistics, National Tsing Hua University, Hsin-chu, Taiwan
[*]Dept. of Electrical Engineering, Chang Gung University, Taoyuan, Taiwan
[†]Dept. of Computer Science and Information Engineering, Chang Gung University, Taiwan
siong@msp.csie.cgu.edu.tw, rylyu@mail.cgu.edu.tw Tel: 886-3-2118800 ext 5709

Abstract

Collection of Taiwanese text corpus with phonetic transcription suffers from the problems of multiple pronunciation, or pronunciation variation. By further augmenting the text with read speech, and using automatic speech recognition with a sausage searching net constructed from the multiple pronunciations of the text corresponding to its speech utterance, we are able to reduce the effort for phonetic transcription. Compared to general method for pronunciation variation such as the relabeling of training corpus of [1], the sausage searching net shows advantages. Two experiments are conducted using a Taiwanese Buddhist Sutra speech and text corpus.

1. Introduction

The majority of the Chinese characters in Taiwanese, as well as other languages in the Han language family such as Hakka and Cantonese, have more than one pronunciation. This is in contrast to the case of Mandarin, where the problem of multiple pronunciations is comparatively much smaller. A Chinese character in Taiwanese commonly can have a classic (Chinese) literature pronunciation (CLP) and a daily life pronunciation (DLP). Many of the characters have even more pronunciations [2]. Although being the mother tongue of more than 70% of the population, Taiwanese was marginalized in Taiwan mainly due to the political suppression in early days. Lack of education in Taiwanese makes the CLP unfamiliar to most people and most people not capable of phonetically transcribing Taiwanese syllables. Besides the multiple pronunciations, Taiwanese also have rich tone sandhis that make it even harder to find person capable of annotating speech data.

In preparing a speech corpus for speech recognition, one usually transcribes the speech data with texts, not with its phonetic transcriptions. Some imperfect automatic transformation from grapheme to phoneme is needed in the training of acoustic models with such a corpus and will inevitably confuse the models. Previous reports on this problem usually come under the title pronunciation variations (see, for example, [1] [3][4]). Taiwanese indeed have the pronunciation variation problem, and some are described in Section 3. We use the term multiple pronunciations to stress the fact that it causes more deteriorations in speech recognition performance.

Buddhism is a major religion in Taiwan (23% of the population, according to [5]), and the Buddhist Sutra, translated into Chinese characters in a terse ancient style (古文), are commonly read in Taiwanese. There are a great number of Sutra volumes. Despite of the popularity, people reading the Sutra tends to make many pronunciation mistakes in characters that require a CLP. It is therefore in dear need for phonetically transcribed versions of the Sutra. Owing to the fact that human expert capable of phonetic transcribing the Sutra in Taiwanese is hard to find, the first volumes of phonetically transcribed Sutra in Taiwanese didn't appear until only in 2004 [6][7], with helps and semi-automatic tools from authors of this paper. Since more transcribed Sutras are planned, we are interested in what automatic speech recognition (ASR) technology can do to the situation. Several attempts have been done in [8][9].

The planned coming Sutra volumes with phonetic transcription are prepared as follows. The Sutra text is segmented into a series of sentences, and each sentence is read and recorded by a senior master nun. It is then ASR's task to phonetic transcribing the text, followed by manual correction from a human expert. The recorded speech will be included for publication in audio CDs. The published Sutra volumes [6] [7] also come with recorded speech in several CDs by the same senior master nun, and thus we may potentially have a speaker dependent ASR to aid the transcription task. Compared to a conventional ASR task, we do know the text associated with the speech. Each character (syllable) in the text has a number of possible pronunciations from a given lexicon, and our task is to discover which of them is actually pronounced.

Note that the procedure just described is more about collecting phonetic annotated text corpus than speech corpus for the training in speech recognition. After all, speaker dependent speech corpus is less useful. Given a set of texts, we are interested in the training of the language model for a speech recognition system. It is much easier to acquire a person to record his/her reading of the text than acquiring a transcribing expert, and thus the effort for the task of phonetic transcriptions can be substantially reduced. For marginalized languages with serious multiple pronunciations problem, this technique is very much welcome.

This paper reports two experiments using the speech and text data in [7] (called TBS corpus for Taiwanese Buddhist Sutra corpus) and is organized as follows. Given speech corpus with phonetic transcription for training, Section 2 reports recognition results on speech with text for its phonetic transcription. Section 3 discusses the second experiment on the recognition of speech without corresponding text under various conditions on the training corpus. Section 4 is a conclusion.

2. First Experiment

The first experiment is on the Sutra transcription problem, and we will discuss its recognition net, the lexicon, the acoustic models and the results.

For multi-syllabic languages such as English, an ASR for dictation machine application requires, among other things, a looped recognition net of all words, equipped with appropriate language model. Its counterpart of a mono-syllabic language such as Taiwanese can be a looped net of all syllables, as in

Fig.1. Such a looped net will be called free-syllable net, if each syllable follows freely, or equally likely, another syllable.

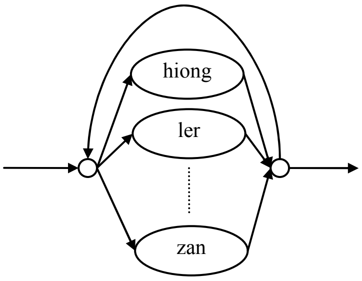

Fig. 1. *Free syllable recognition net.*

For our Sutra transcription problem, we have, in addition to each speech utterance, its associated text in Chinese character. Based on the multiple pronunciations of each Chinese character, we can construct a much smaller recognition net. An example for the utterance "爐香讚" is in Fig. 2. We will call such a net (multiple pronunciations) sausage net for its shape, following [10]. Higher recognition results can be expected due to smaller complexity in the recognition net. Our task is then amount to how to construct sausage nets and which acoustic model to choose.

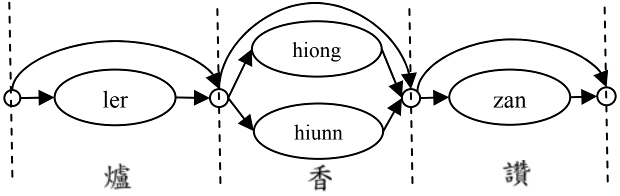

Fig. 2. *The sausage searching net. The net is constructed from the multiple pronunciations of each Chinese character from our Formosa Lexicons. The corresponding Chinese characters of multiple pronunciations are also shown.*

2.1. The pronunciation Lexicons

There are three pronunciation lexicons for the multiple pronunciations in Taiwanese of the Chinese characters.

The first is **Formosa Lexicon**. It is a combination of two lexicons: Formosa Mandarin-Taiwanese Bi-lingual lexicon and Gang's Taiwanese lexicon [10][11][12]. The former is derived from Mandarin lexicon, and thus many commonly used Taiwanese terms are missing. The latter contains Taiwanese expressions from a sampling of radio talk show. Some statistics of the two sub-lexicons are summarized in Table 1 and 2.

	CLP Taiwanese	DLP Taiwanese	Total
1-Syllable	2319	8040	10359
2-Syllable	21337	49222	70559
3-Syllable	7163	11367	18530
4-Syllable	55	15525	15580
5-Syllable	1	711	712
6-Syllable	0	497	497
7-Syllable	0	478	478
8-Syllable	0	195	195
9-Syllable	0	3	3
10-Syllable	0	20	20
Total	30875	86060	116935

Table 1. *The number of pronunciation of Formosa bi-lingual Lexicons, including classic literature pronunciation (CLP) and daily life pronunciation (DLP).*

	Taiwanese
1-Syllable	8153
2- Syllable	46587
3- Syllable	13241
4- Syllable	2106
5- Syllable	175
Total	70262

Table 2. *The distribution of Gang's Taiwanese lexicon.*

Sutra Lexicon. The second lexicon is the Sutra derived lexicon. It is the pronunciations collected from the published volumes of the Sutra. (Only those in [7] are included for this experiment.) The general lexicon, used for wider range of applications, tends to have higher number of multiple pronunciations. We expect this lexicon to have a high "hit rate."

The third lexicon is the combination of the previous two, and called **Enhanced Lexicon**.

2.2. The Recognition Net

For the searching net of the ASR, we have four choices.

The first is the free-syllable net. It is simply a looped net of all Taiwanese syllables in TBS, which the number of syllable is 467, denoted as F-Syl Net.

The other three searching nets are the sausage nets generated from each of the three pronunciation lexicons. The nets are denoted as FL-S Net, SL-S Net, and EL-S Net for the Formosa, Sutra, and Enhanced Lexicon respectively.

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 a problem since there are simply too many Chinese characters. With Unicode Standard contains more than thirty thousands Chinese characters, the Sutra still have some not in the Unicode character set. The Formosa Lexicon has much less distinct characters, and the missing characters problem is inevitable. When a missing character is encountered, we use all possible syllables as its multiple pronunciations. Fig. 3 illustrates this case.

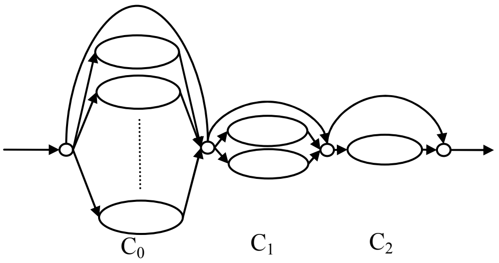

Fig. 3. *The sausage searching net with missing character C_0 assuming all syllables as its possible pronunciations.*

The missing pronunciation problem is one that the real pronunciation of a character is not included in the lexicon, although it has entry for that character with different pronunciation. Since each character could have potentially missing pronunciation in the lexicon, we do nothing to prevent the problem. It is thus the task of human expert to correct the missing pronunciation in the later stage.

2.3. The Acoustic Models

For the acoustic model, we can choose speaker dependent or speaker independent model.

The speaker independent model is trained from our ForSDAT bilingual (Taiwanese and Mandarin) speech corpus.

	Name	Gender	Quantity	Time(hr)
ForSDAT	TW01-M0	Female	50	6.21
	TW01-M1	Male	50	5.44
	MD01-M0	Female	50	5.92
	MD01-M1	Male	50	5.42
	TW02-M0	Female	233	10.80
	TW02-M1	Male	277	11.66

Table 3. *The ForSDAT microphone speech corpus.*

The TW02 set is not used for acoustic training. It is used for pronunciation variation rules as discussed in next subsection.

Speaker dependent model is possible for this experiment because the Sutra speech is recorded by the same master nun. The speech data [7] has 3,149 utterances in 350 minutes. 160/31 utterances are randomly chosen and reserved for testing and acoustic model development (speaker adaptation).

Buddhist Corpus Category	Utterance	Time(min)
Train	2958	333.87
Development	31	2.56
Test	160	13.33
Total	3149	349.76

Table 4. *TBS (Taiwanese Buddhist Sutra) speech corpus.*

2.4. The Result

All the speech data are recorded as microphone speech with 16K, 16bits PCM speech data. For the acoustic model, we use continuous Gaussian-mixture HMM model with feature 52 dimension MFCC computed using 20ms window frame and 10ms frame shift. Context-dependent inside-syllable tri-phone models were built using a decision-tree state tying procedure. Most of the training and recognition are carried out using the HTK tool.

With the four searching net and three acoustic models, table 5 shows the recognition results.

Acoustic model	Searching net	Accuracy (%)
SI w/o adaptation	F-Syl Net	25.34
	FL-S Net	81.80
	SL-S Net	90.21
	EL-S Net	84.66
SI w/ adaptation	F-Syl Net	55.34
	FL-S Net	84.55
	SL-S Net	91.48
	EL-S Net	88.31
SD	F-Syl Net	82.49
	FL-S Net	87.74
	SL-S Net	91.75
	EL-S Net	91.84

Table 5. *Recognition results under four searching nets. See text in section 2.2 for notations.*

With sufficient data, SD model should perform better than SI model. However, under the multiple pronunciation sausage searching net, the SI model can compete with the SD model. Being able to use the speaker independent model for phonetic annotation task is very much welcomed.

3. Second Experiment

3.1. The problem

In the previous section, the training of the speaker dependent model uses text data with phonetic transcription prepared by human expert. What if only the text is available, not its phonetic transcription? This type of situation happens for all languages since words commonly have pronunciation variations, and preparing a "correct" phonetic transcription for the training is not a trivial task. For Taiwanese, it is more serious: the multiple pronunciations of a Chinese character is not just minor variations in pronunciation, they are usually completely different syllables. We are thus interested in how the ASR performs in this case.

Basically this is an ASR-aided phonetic annotation problem for the speech data. It happens in our effort of collecting annotated Taiwanese text corpus. We could have a person (thus speaker dependent) read along a Taiwanese text and record, and then use ASR helping the phonetic annotation. The TBS corpus provides us a good setup for simulating such a process.

3.2. The Procedure

The procedure goes as follows. First an automatic procedure is applied to annotate the training corpus, conduct the training of the speaker dependent model with the TBS corpus while treating the annotations as if it is true, and then recognize the testing utterances. There are four automatic annotation approaches are used here.

(1). G2P (Grapheme-to-Phoneme). For each text in sentence, this is simply maximal matching the text against the Formosa Lexicon. In case of multiple pronunciation of a word or Chinese character, choose one arbitrary. Note that the annotation involves no ASR for G2P.

(2). SI/FL-S-Net. The SI model with speaker adaptation as in last section is used here to annotate the training data in conjunction with FL-S-Net.

(3). SI/EL-S-Net. Similar to procedure (2), but the searching net is EL-S-Net.

(4). SI/AL-S-Net. Similar to procedure (2), but the searching sausage net is generated by an adapted version of the Formosa lexicon. Explanation follows.

With Taiwanese being marginalized, some people, especially younger ones, make pronunciation mistakes when speak in Taiwanese. In our partially manually validated ForSDAT TW02 speech corpus, we observed many such mistakes. Table 6 shows another set of some well-known pronunciation mistakes. Table 7 shows some of the most frequent substitution errors.

	Pattern	Pattern
Knowledge-based rules	zh → z	n → l
	ch → c	l → n
	sh → s	er → o
	rh → r	r → l
	n → l	gh → g
	bh → m	gh → {}
	m → bh	h → {}
	bh → b	

Table 6. *The knowledge-based error patterns.*

Applying the rules to each syllable in the lexicon, we virtually end up with much larger number of multiple pronunciations for each Chinese character. With this adaptation in the lexicon, we call the sausage net so generated AL-S-Net.

	Error Pattern	Error Count
Statistics-based rules	i-ng → i-n	126
	bh-er → bh-o	50
	i-m → i-n	44
	a-m → a-n	43
	inn-onn → inn-unn	29
	a-m → a-ng	26
	a-ng → a-m	25
	a-n → a-m	23
	i-m → i-ng	21
	b-er → b-o	19
	i-n → i-m	19
	d-er → d-o	18
	i-a+ng → i-o+ng	17

Table 7. *The 13 most frequent substitution errors from the partially validated ForSDAT-TW02 corpus.*

3.3. The Result

The result of the automatic annotation is, of course, not perfect. But with these annotations, we proceed to train speaker dependent models using the TBS speech corpus, and then perform the recognition on the 160 testing utterances with free syllable net. The results are in Table 8.

Training Text Annotation Method	Accuracy rate for TBS testing set
(1) G2P[I]	75.56%
(2) SI/FL-S-Net	77.30%
(3) SI/EL-S-Net	80.16%
(4) SI/AL-S-Net	72.96%
* Manual annotated training corpus	82.49%
* Speaker independent model with adaptation (Sec 2.4)	55.34%

Table 8. *Recognition results applying the four annotation approaches to the training corpus. See text for explanation of notations. The fifth row is for that using manually annotated training corpus.*

We include the result from that using manual annotation in the 5th row for comparison. It should be the upper limit. From Table 8 it appears that every approach leads to satisfactory result, considering we are not using more powerful language model such as tri-gram. Also, when compared to the speaker independent model, it clearly shows the advantages of speaker dependent model. It also shows that restricting the search net to sausage net does overcome the advantages of speaker dependent model, as exhibited in the last section. Note that we do not use the sausage net in this experiment to show a welcome partial solution to the problem of not having phonetic transcription in the training corpus.

4. Conclusions

Taiwanese text corpus collection suffers from the multiple pronunciation problems. By further augmenting the text with read speech, we are able to reduce the effort for phonetic transcription of the text using automatic speech recognition with a sausage searching net constructed from the text corresponding to its speech utterance. Our approach shows advantages over, say, the method of re-labeling the training corpus of [1] which does not utilize the special multiple pronunciations.

5. References

[1] Kanokphara, Supphanat, Virongrong Tesprasit and Rachod Thongprasirt, “Pronunciation Variation Speech Recognition without Dictionary Modification on Sparse Database”, In Proc. *ICASSP 2003*, Hong Kong, 2003.

[2] Liang, Min-siong, Ren-yuan Lyu, Yuang-chin Chiang, and Dau-Cheng Lyu, “A Taiwanese Text-to-Speech System with Applications to Language Learning,” In Proc. *ICALT*, Joensuu, Finland, 2004.

[3] Tsai, Ming-Yi, Fu-chiang Chou, and Lin-shan Lee, “Improved pronunciation modeling by inverse word frequency and pronunciation entropy”, IEEE ASRU 2002, pp. 53-56.

[4] Raux, Antoine, “Automated Lexical Adaptation and Speaker Clustering based on Pronunciation Habits for Non-Native Speech Recognition”, In Proc. *International Conference on Spoken Language Processing*, Jeju Island, Korea, 2004.

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

[6] Sik, DatGuan, *The Four Basic Sutra in Taiwanese*, DiGuan Temple, HsinChu, Taiwan, 2004.

[7] Sik, DatGuan, *Earth Store Sutra in Taiwanese*, DiGuan Temple, HsinChu, Taiwan, 2004.

[8] Siohan, Olivier, Bhuvana Ramabhadran and Geoffrey Zweig, “Speech Recognition Error Analysis on the English MALACH Corpus”, In Proc. *ICSLP 2004*, Jeju Island, Korea, 2004.

[9] Ramabhadran, Bhuvana, Jing Huang and Michael Picheny, “Towards Automatic Transcription of Large Spoken Archives - English ASR the MALACH Project”, In Proc. *ICASSP*, Hong Kong, China, 2003.

[10] Mangu, Lidia, Eric Brill and Andreas Stolcke, "Finding Consensus in Speech Recognition: word error minimization and other applications of confusion networks", Computer Speech and Language, 2000.

[11] Liang, Min-Siong, Ren-Yuan Lyu, Yuang-Chin Chiang, “An Efficient Algorithm to Select Phonetically Balanced Scripts for Constructing A speech Corpus”, *IEEE-NLPKE 2003*, October 26-29, 2003, Beijing, China.

[12] Liang, Min-siong Dau-cheng Lyu, Yuang-chin Chiang, Ren-yuan Lyu, “Construct a Multi-Lingual Speech Corpus in Taiwan with Extracting Phonetically Balanced Articles”, In Proc. *ICSLP 2004*, Jeju Island, Korea, 2004.

[13] Lyu, Ren-yuan et al., “Toward Constructing A Multilingual Speech Corpus for Taiwanese (Minnan), Hakka, and Mandarin”, *ICLCLP Vol. 9, No. 2, August 2004 pp. 1-12*