Using Speech Recognition Technique for Constructing a Phonetically Transcribed Taiwanese (Min-nan) Text Corpus

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

*Dept. of Electrical Engineering, Chang Gung University, Taoyuan, Taiwan
Dept. of Computer Science and Information Engineering, Chang Gung University, Taiwan
+ Inst. of Statistics, National Tsing Hua University, Hsin-chu, Taiwan
{minsiong, renyuan.lyu}@gmail.com, Tel: 886-3-2118800 ext 5967, Fax: 886-3-2118490

Abstract

Collection of Taiwanese text corpus with phonetic transcription suffers from the problems of multiple pronunciation variation. By augmenting the text with 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. By using the multiple pronunciation lexicon, the error rate of transcription 13.94% was achieved. Further improvement can be achieved by adapting the pronunciation lexicon with pronunciation variation (PV) rules derived from a manual corrected speech corpus. The PV rules can be categorized into two kinds: the knowledge-based and data-driven rules. By incorporating the PV rules, the error rate reduction 13.63% could be achieved. Although the technique was developed for Taiwanese speech, it could also be adapted easily to be applied in the other similar "minority" Chinese spoken languages.

1. Introduction

Automatically phonetic transcription of text is gaining popularity recently in speech processing field, especially in speech recognition, text-to-speech and speech database construction [1]. The problem may become rather non-trivial when the target text is the Chinese text (漢字). The Chinese writing system is widely used in China and the East/South Asian countries/areas including Taiwan, Singapore, and Hong-kong. It is also adopted historically in Japan, Korea and Vietnam. It is known (and pronounced) as "hànzì" in Mandarin, "kanji" in Japanese, "hanja" in Korean, and "hànli" in Taiwanese (Min-nan). Fundamental elements in this writing system have been called characters, pictographs, pictograms, ideograms, etc. In this paper, they shall be called hanzi or Chinese characters for simplicity and clarity. Although the same hanzi is used in different countries/areas, the pronunciation may be very different.

Taiwanese Buddhism Sutra (written collections of Buddhist's teaching in Min-nan language) was chosen to be the target of text corpus to be processed in this paper. Unlike the other major Chinese spoken languages, such as Mandarin or Cantonese, the use of Taiwanese has been banned officially in schools, radio and TV programs for a long time during the past decades. More and more transcribed text is then needed to teach children and adults to read the Chinese text in Taiwanese speech. Although many grapheme-to-phoneme systems exist, almost none of them could be used directly in this task. The reasons include: no qualified pronunciation lexicon exist, very few appropriately computational linguistic research were conducted to support such a grapheme-to-phoneme system. Other reasons are due to the severe multiple variations for Taiwanese speech. This phenomenon motivate us to develop an automatic phonetic transcription technique or system to transcribe the large quantity of Chinese text into Taiwanese phonetic symbols such that these texts could be used as Taiwanese language read-aloud tutorial materials. The case becomes even more difficult when considering the complex tone sandhi phenomenon in Taiwanese.

The majority of the Chinese characters in Taiwanese, as well as other languages in the Chinese language family such as Hakka and Cantonese, have more than one pronunciations. This is in contrast to the case of Mandarin, where the problem of multiple pronunciations is less important. A Chinese character in Taiwanese commonly can 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) [2]. In additional to the problem of multiple pronunciations due to the variation of Wen-du-in and Bai-du-in, Taiwanese also has the pronunciation variation due to the sub-dialectical accents, such as Tainan-, Taipei-, and Lugang- accents. Previous reports on this problem usually come under the title pronunciation variations [3]. We use the term multiple pronunciations to stress the fact that it may cause more deteriorations for phonetic transcription.

As shown in Fig. 1, the Sutra text is segmented into a series of sentences, and each sentence is read and recorded by a senior master nun. Then ASR technique is applied to phonetically transcribe the text, followed by manual correction from a human expert. By sending the corrected transcription feedback into ASR system, the accuracy could be improved until an acceptable performance could be achieved. 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.

2. The Phonetic Transcription using Speech Recognition Technique

The whole framework can be divided into two major parts, i.e. an acoustic part and a language part. Thus we define: S is the syllable sequence, while O and W are the observed acoustic sequence and the input character sequence. The phonetic transcription target is to find the most possible syllable sequence S^* given O and W. The formula is:

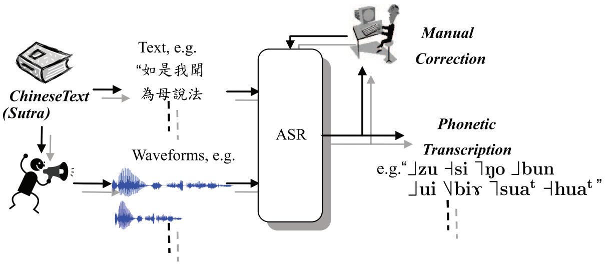

Figure 1: *The process of semi-automatic transcription of Chinese text to Taiwanese pronunciation using ASR technique.*

$$S^* = \arg\max_S P(S|O, W) \tag{1}$$

By the Bayes equation:

$$\begin{aligned} S^* &= \arg\max_S \frac{P(S|W)P(O|S,W)}{P(O|W)} \\ &= \arg\max_S P(S|W)P(O|S,W) \end{aligned} \tag{2}$$

Assume $W \subset S$, Eq. 2 is simplified as:

$$S^* = \arg\max_S P(S|W)P(O|S) \tag{3}$$

The first part of Eq. 3 is independent of O and defined as language model in recognition. The second part is the probability of observation given the syllable sequence and defined as acoustic part.

For the acoustic part, we can choose speaker dependent (SD) or speaker independent model (SI). The speaker independent model is trained from the ForSDAT-01 bilingual (Taiwanese and Mandarin) speech corpus, which contains 200 speakers and 23 hours of speech, including both Taiwanese and Mandarin. All the speech data are recorded in 16K, 16bits PCM format. We use continuous Gaussian-mixture HMM models with feature vectors of 52-dimension MFCC computed by using 20-ms window frame and 10-ms frame shift. Context-dependent inside-syllabic tri-phone models were built using a decision-tree state tying procedure.

For the acoustic part, we can choose speaker dependent (SD) or speaker independent model (SI). The speaker independent model is trained from the ForSDAT-01 bilingual (Taiwanese and Mandarin) speech corpus, which contains 200 speakers and 23 hours of speech, including both Taiwanese and Mandarin. All the speech data are recorded in 16K, 16bits PCM format. We use continuous Gaussian-mixture HMM models with feature vectors of 52-dimension MFCC computed by using 20-ms window frame and 10-ms frame shift. Context-dependent inside-syllabic tri-phone models were built using a decision-tree state tying procedure.

In the task considered here, it is possible to use a set of SD models for the experiment because the quantity of Sutra speech recorded by the same master nun is quite large. The distribution of Sutra speech is listed in Table 1. 160 utterances are randomly chosen and reserved for testing while as another 31 utterances were used for acoustic model development. Maximum Likelihood Linear Regression (MLLR) is then used to adapt speaker independent models.

For the language part, the problem of multiple pronunciations could be solved by using specially designed searching net. All the searching nets were constructed according to multiple pronunciation lexicon described in next section. Finally, the pronunciation variation rules would be incorporated in searching net to improve the accuracy of transcription as discussed in the section 4.

Table 1: *TBS (Taiwanese Buddhist Sutra) speech corpus.*

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

3. Baseline Experiments in Sausage Network

For the Sutra transcription problem, in addition to each speech utterance, its associated text in form of Chinese characters was also another input. Assume all syllables are independent with each other. Therefore, the Eq. 3 can be rewritten as:

$$S^* = \arg\max_S P(s_1|w_1)...P(s_n|w_n)P(O|s_1, s_2, ..., s_n) \tag{4}$$

Here we encounter two problems: one is what syllable could be the pronunciation for the Chinese character, the other is what probability should be given for the pronunciation. By looking the pronunciation up in pronunciation lexicon, the multiple pronunciations of each Chinese character could be found. Based on the multiple pronunciations of each Chinese character, a much smaller recognition net can be constructed. We will call such a net (with multiple pronunciations) as a “sausage” net for its shape. Higher recognition accuracy can be expected due to its smaller complexity in the recognition net. Our task is then amount to how to construct sausage nets and which acoustic model to choose.

3.1. The pronunciation Lexicons and the Recognition Net

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

The first is **Formosa Lexicon**, which contains about 123 thousand words in Chinese/Taiwanese text with Mandarin/Taiwanese pronunciations. It is a combination of two lexicons: Formosa Mandarin-Taiwanese Bi-lingual lexicon and Gang’s Taiwanese lexicon [4]. The former is derived from a Mandarin lexicon, and thus many commonly used Taiwanese terms are missing due to the fundamental difference between these two languages. The Formosa 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. But some pronunciations, which are actually pronounced by some experts, do not appear in Formosa Lexicon due to pronunciation variations. Sometimes, the Formosa Lexicon does not contain some ancient characters of Sutra. Thus, the second lexicon, called **Sutra Lexicon**, is derived from the Sutra itself to study what variations exist from Formosa lexicon to Sutra Lexicon. It is the pronunciations collected from the published volumes of the Sutra. We expect this lexicon to cover those words/characters/pronunciations which were not contained in the Formosa Lexicon mentioned above. The performance of Sutra Lexicon would be expected the best result or the upper bound for the phonetic transcription considered here. The third lexicon is not another source of pronunciation but just the combination (union) of the previous two, and called **Enhanced Lexicon** for convenience.

The three searching nets are the sausage nets generated from each of the three pronunciation lexicons. Each searching net was constructed by filling in each node of the net with the corresponding multiple pronunciations of each Chinese character from the pronunciation lexicon. The nets are denoted as General-Sau-Net, Specific-Sau-Net, and Enhanced-Sau-Net for the general-purpose Formosa Lexicon, the specific-domain Sutra Lexicon, and the combined Enhanced Lexicon, respectively.

3.2. The Recognition Results

With the three searching nets (General-Sau-Net, Specific-Sau-Net, and Enhanced-Sau-Net) and three acoustic models (SI with adaptation, SI without adaptation, and SD), the recognition results measured as syllable error rate (SER) are shown in Fig. 2.

The transcription by using unreliable SI model may lead worse result from observing the results 21.31% and 24.02% error rates corresponding to General-Sau-Net and Enhanced-Sau-Net with SI model, even though the Enhanced-Sau-Net covers more possible pronunciations than General-Sau-Net. On the other hand, whatever the nets were used, the performance of Specific-Sau-Net was invariably the best among those nets. However, General-Sau-Net could surpass Free-Syl-Net, and compete with Specific-Sau-Net. For example, under the same speaker adaptation models, the result 13.94% with General-Sau-Net was better than the result 43.85% with Free-Syl-Net and moderate worse than 8.64% with Specific-Sau-Net.

If the sufficient data was available for speaker dependent model under the Specific-Sau-Net, the minimum error rate 4.98% could be achieved. In practice, this environment for recognition was not easy to construct. Thus, if speaker independent model could be adapted by a little phonetically transcribed speech data, the adapted speaker independent model under General-Sau-Net for phonetic annotation task is very much welcomed and the result 13.94% could be the baseline experiment.

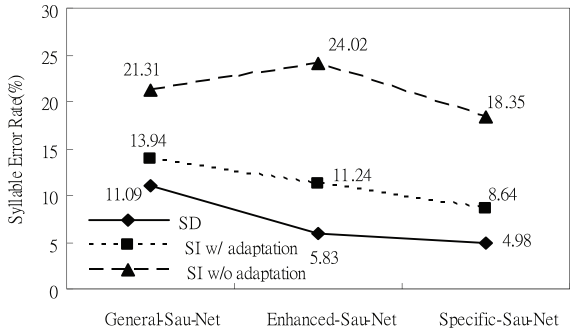

Figure 2: *Syllable error rate (SER) under four searching nets. See text in subsection 3.1 for notations.*

4. Incorporating Pronunciation Variation Rules

Because insufficient coverage of pronunciations in searching net will degrade the recognition performance severely, some approaches to extend the coverage of pronunciations will be considered to help the overall performance. Some rule-derived variant pronunciations are added in the searching net directly to enhance the pronunciation coverage. Generally speaking, the pronunciation-variation (PV) rules can be categorized into two kinds: the knowledge-based and the data-driven rules. The knowledge-based rules were derived from the knowledge established by phonetician experts. On the other hand, the availability of transcribed real speech corpus will provide alternative source to help derive some data-driven PV rules by statistical methods.

4.1. The knowledge-based variation rules

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 the following points:

1) Variation between ***Bai-du-in*** and ***Wen-du-in***: The variations may vary due to classic literate pronunciation (known as Wen-du-in) and a colloquial pronunciation (known as Bai-du-in),which were 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.

2) Variation between sub-dialectal regions: Some variations were referred to dialectal differences; For instance, the initials /z/ is substituted with /l/ or /ɡ/ depending on the sub-dialects of Taiwanese, where the phenomenon was denoted as “z→l/ɡ”.

3) Variation due to personal pronunciation errors: Some kinds of variations are considered as personal pronunciation errors. For example, some pronunciation may disappear in younger generation such as phoneme /ɡ/, where the phenomenon is denoted as “ɡ→{}”.

The knowledge source for the pronunciation variation rule selection. But the knowledge-based PV rules, which were derived by more than one linguists, were sometimes contradictory with each other. This made them sometimes hard to choose in implementation. Of course, some of pronunciation variation rules are certainly language-dependent (i.e. the phonological and phonetic processes differ between languages). However, the major points to be emphasized are that the proposed technique to model pronunciation variation for transcription was rather language-independent.

4.2. The data-driven variation rules

The form $LBR \rightarrow LSR$ (triphone) is suitable for representing 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 context respectively. The number of triphone units in Taiwanese is about 1200. To derive such rules, a speech corpus with both canonical pronunciation and actual pronunciation is necessary. The Formosa Speech Database (ForSDAT) is such a database, which is collected under the grant of National Science Council of Taiwan during the past several years [4]. A subset of ForSDAT, called ForSDAT-TW02, was used in this project.

ForSDAT-TW02 is a bi-phone rich speech database. A small portion of the speech data has then been manually checked and the phonetic transcription of the transcript is “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 2, which is called the confusion tables in syllable level and triphone level.

Three kinds of statistical measures were used in this paper. They are (1) Joint probability, (2) Conditional probability, and (3) Mutual information of the base form pronunciation and the surface form pronunciation. The mathematic definitions of the above 3 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, conditioning on the base form pronunciation b_i,$p(s_j|b_i) = n_{ij}/N_i$

(3)Mutual information of the base form pronunciation b_i, and the surface form pronunciation s_j,

$$I_{ij} = p(b_i, s_j) log\frac{p(b_i,s_j)}{p(b_i)p(s_j)} = \frac{n_{ij}}{N_i} log(N \cdot \frac{n_{ij}}{\sum_i n_{ij} \cdot \sum_j n_{ij}})$$

In all the above equations, n_{ij} is the number of substitutions of (base-form) triphone bi by the surface-form triphone s_j that appear in a corpus, where $N = \sum_i \sum_j n_{ij}$, $N_i = \sum_j n_{ij}$. While $p(b_i, s_j)$ represents the joint probability of (b_i, s_j), and $p(b_i)$, $p(s_j)$ equal the marginal probability of b_i and s_j.

Table 2: Triphone-level confusion table, where n_{ij} represent the number of variation from triphone bi 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 $N = \sum_i \sum_j n_{ij}$.

	b-ɤ	i-n	…	s_j	…	b-o	
b-ɤ	237	0	…	n_{1i}	…	30	267
i-ŋ	0	84	…	n_{2i}	…	0	1373
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
b_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{iP}	N_i
	241	1102	…	M_j	…	107	N

4.3. The Recognition Results

We adapt the Formosa (general-purpose) pronunciation lexicon according to different sets of pronunciation variation rules. Then the speech recognition task with sausage net and speaker adaptation was conducted as described in section 3, where the SER achieved before the application of the pronunciation variation rules was 13.94% as shown in Fig. 2, and would be looked upon as the performance of the baseline setup in this section. In Fig. 3, we could observe that it is truly helpful to decrease the SER by increasing the coverage of searching net via the usage of PV rules. The evidence is that the lowest error rate 12.04% is achieved by utilizing the first 52 variation rules selected by Mutual-Information (MI) method. Similar improvement would also be observed in the best SER 12.88% and 12.51% achieved by Joint-Probability (JP) method and Conditional-Probability (CP) method.

After applying more rules, the SER does increase. The situation will even become worse than the baseline experiments. Therefore, it is an important way that how to gather the better rules into the preceding rules and let the worse rules appear latter.

Although the JP-based method could make the error rate converge more quickly than CP-based method, the performance also degraded mostly quickly. This is because the score of CP-based method would be normalized by his base-form count in contrast to JP-based method. But sometimes few count of the PV-rule might get the higher conditional probability due to few observations of the base-form. Therefore, many insignificant and harmless PV-rules of CP-based method were appeared in high rank and lead the slowest convergence among these three methods.

In MI-based method, the formula could avoid the slow convergence by using the Joint-Probability as weight since few count of the base-form would get few numbers of variations. From 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 the mutual information between base-form and surface-form to calculate the correlation of base-form and surface-form by the normalization of their count. Consequently, the error rate of the MI rank converges most quickly and the performance of MI method in error reduction is also better than JP method and CP method respectively.

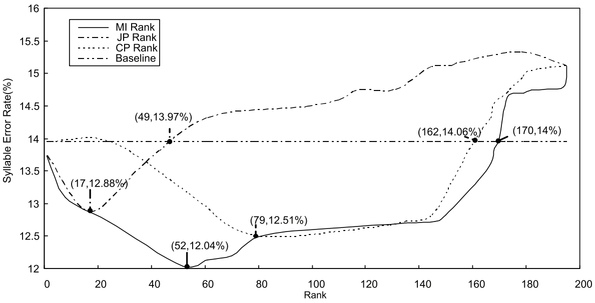

Figure 3: *The recognition result (syllable error rate) v.s. the number of ranked rules sorted according to different measures, including mutual-information (MI), joint-probability (JP), and conditional-probability (CP) as well as the Baseline criterion.*

5. Conclusion

We have proposed a new approach to address the phonetic transcription of Chinese text into Taiwanese pronunciation. The proposed semi-automatic transcription of Chinese text to Taiwanese pronunciation system has reached the 13.94% error rate as baseline experiment. The further improvement by using pronunciation variation rules had 13.63% error rate reduction.

6. References

[1] Soltau, H., "The IBM 2004 Conversational Telephony System for Rich Transcription", In: Proc. ICASSP, Philadelphia, USA, 2005, pp. I-205-I-208.

[2] Liang, M.-S., et al., "A Taiwanese Text-to-Speech System with Applications to Language Learning", In: Proc. ICALT, Joensuu, Finland, 2004 pp. 91-95.

[3] Hain, T., "Implicit modelling of pronunciation variation in automatic speech recognition", Speech Communication 46, 2005, pp. 171-188.

[4] Lyu, Ren-yuan et al., "Toward Constructing A Multilingual Speech Corpus for Taiwanese (Minnan), Hakka, and Mandarin", International Journal of Computational Linguistics & Chinese Language Processing (IJCLCLP), Vol. 9, No. 2, August 2004, pp. 1-12.