PHONETIC TRANSCRIPTION USING SPEECH RECOGNITION TECHNIQUE CONSIDERING VARIATIONS IN PRONUNCIATION

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

Dept. of Electrical Engineering, Chang Gung University, Taiwan.
Dept. of Computer Science and Information Engineering, Chang Gung University, Taiwan.
Institute of Statistics, National Tsing Hua University, Taiwan.
*E-mail: renyuan.lyu@gmail.com

ABSTRACT

We propose a new approach for performing phonetic transcription of speech and text that combines automatic speech recognition (ASR) and grapheme -to- phoneme (G2P) techniques. By augmenting the text with speech and using automatic speech recognition with a sausage searching net constructed from multiple text pronunciations corresponding to human speech utterance, we are able to reduce the effort for phonetic transcription. By using a multiple pronunciation lexicon, a transcription error rate of 12.74% was achieved. Further improvement can be achieved by adapting the pronunciation lexicon with pronunciation variation (PV) rules and an error rate reduction of 17.11% could be achieved.

Index Terms— Automatic Phonetic Transcription, Pronunciation Variation, Chinese, Taiwanese, Dialect.

1. INTRODUCTION

Automatic phonetic transcription is gaining popularity in the speech processing field, especially in speech recognition, text-to-speech and speech database construction [1]. It is traditionally performed using two different approaches: an acoustic feature input method and text input method. The former is the speech recognition task, or more specifically, the phoneme recognition task. The latter is the 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 away 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 the 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 ***a text corpus*** derived from the ***Buddhist Sutra*** (written collections of Buddhist teachings). Buddhism is a major religion in Taiwan (23% of the population). The Buddhist Sutra, translated into Chinese text in a terse ancient style (古文) , is commonly read in ***Taiwanese (Min-nan)***. Due to 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 researches were conducted to support developing a G2P system.

Taiwanese uses Chinese characters as a part of the written form, with its own phonetic system, which is very different 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 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 addition to MPs, Taiwanese also have a ***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.

The traditional approach to transcribing Chinese Buddhist Sutra text uses human dictation. A master monk or nun reads the text aloud, sentence by sentence. The manual transcription process is tedious and prone to errors. Since more transcribed Sutras are planned, we are interested in how ASR and G2P technology can help in this situation. Our task is to discover which of them is actually pronounced. It is much easier to acquire a person 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.

2. THE PHONETIC TRANSCRIPTION TECHNIQUE

The flow chart shown in Fig. 1 is the framework of phonetic transcription using the speech recognition technique. Based on flow chart in Fig. 1, we define: $\underline{s}$ is the syllable sequence,

Table 1. The statistics of total ForSDAT-01 speech corpus and partially manually validated ForSDAT-02 speech corpus, where M and F denote male and female

	ForSDAT-01	ForSDAT-02
Utterance	92158	19731
People	100(M: 50, F: 50)	131(M: 72, F: 59)
# of Syllables	179730	45865
# of distinct triphones	1356	1194
# of total triphones	555731	104894
Time(hr)	22.43	7.2

while $\underline{c}$ and $\underline{o}$ are the input character and augmented acoustic sequences. The phonetic transcription target finds the most probable syllable sequence $\underline{s}^*$ given $\underline{o}$ and $\underline{c}$. The formula is: $\underline{s}^* = \arg\max_s P(\underline{s}|\underline{o},\underline{c})$, where $\underline{c} \in \underline{C} = \{\underline{c}|\underline{c} = c_1^M = c_1...c_M, c_i \in C\}$, c_i is an arbitrary Chinese character C is the set of all Chinese characters and $\underline{s} \in \underline{S} = \{\underline{s}|\underline{s} = s_1^N = s_1...s_N, s_i \in S\}$, s_i is an arbitrary Taiwanese syllables, S is the set of all Taiwanese syllables. Using the Bayes theorem: $\underline{s}^* = \arg\max_s P(\underline{s}|\underline{c})P(\underline{o}|\underline{s},\underline{c})$.

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

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

The first term, $P(\underline{s}|\underline{c})$, of Eq. 1 is independent of $\underline{o}$ and plays the major role in the language part of the recognition scheme. The second term, $P(\underline{o}|\underline{s})$, is the probability of observation given the syllable sequence and plays the major role in the acoustic part.

For the acoustic part, we choose SI-HMM model trained by Formosa Speech Database 01 (ForSDAT-01) [2]. The features were extracted into vectors of 48 dimensional MFCC plus 4 dimensional energy. Context-dependent tri-phone models were built using a decision-tree state tying procedure. Maximum Likelihood Linear Regression (MLLR) is then used to adapt SI models using 31-utterance Taiwanese Buddhist Sutra (TBS) adaptation data[4].

For the language part, it could be modeled as a traditional G2P problem. Even the best pronunciation lexicon would miss the true pronunciation for a certain Chinese character. To address this issue, the pronunciation variation rules would be incorporated in a sausage net to improve the accuracy of transcription. The rule set of pronunciation variations can be trained from the partial ForSDAT-02 speech corpus. The statistical information of ForsDAT and TBS was summarized as in Table 1 and 2.

3. BASELINE EXPERIMENTS IN SAUSAGE NETWORK

The first experiment is performed on the sausage recognition network without considering the pronunciation variation

Table 2. TBS (Taiwanese Buddhist Sutra) speech corpus.

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

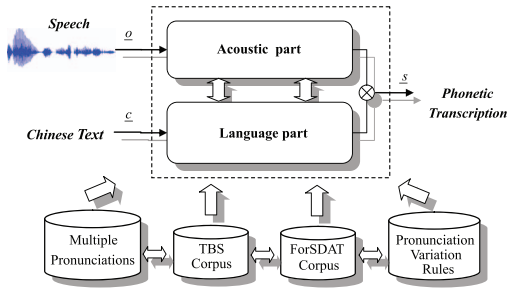

Fig. 1. The flow chart of the phonetic transcription of TBS incorporating pronunciation variation rules.

problem. For a syllabic language, we can construct a looped net of all syllables, called a free-syllable net. Based on the Eq. 1, we define: $\underline{s} = s_1, s_2, ..., s_N$ is the syllable sequence. Assume the character sequence $\underline{c}$ is unknown. All syllables are independent from each other and $P(s_i)$ is a uniform distribution. Following Eq. 1, we have:

$$\underline{s}^* = \arg\max_{\forall \underline{s} \in \underline{S}} P(\underline{o}|s_1, s_2, ..., s_N) \quad (2)$$

It is a compact representation for the search 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.

Secondly, it is also possible to input only text. Because only a small database scale is available, we assume that s_i is dependent only on c_i. In this case, $P_{\underline{S}|\underline{C}}(\underline{s}|\underline{c})$ can be simplified as $\prod_{i=1}^{N} P_{S_i|C_i}(s_i|c_i)$. If there are two knowledge sources will be used to provide the text-pronunciation correspondence information, where one is the pronunciation lexicon L_1 with multiple pronunciations for each Chinese character and the other is a small text corpus L_2, phonetically transcribed by human experts. Then, the term $P_{S_i|C_i}(s_i|c_i)$ can be further expanded, then Eq. 1 becomes

$$\underline{s}^* = \arg\max_{\forall s_i \in S_i} \prod_{i=1}^{N} \sum_{j=1}^{2} P_{S_i|C_i,L}(s_i|c_i, l_j) P_L(l_j) \quad (3)$$

where l represents elements in the set representing the knowledge source, and we let $P(l_1) = P(l_2) = 0.5$ without optimizing $P(l_1)$ and $P(l_2)$. The pronunciation probability in the lexicon is supposed to be a uniform distribution. In addition, $P_{S_i|C_i,L}(s_i|c_i, l_2)$ can be estimated from the small transcribed text corpus. The results from the experiments con-

ducted using Eq. 3 depends only on the text input and are referred as the language part performance.

What is proposed in this paper is an approach to integrate both. Given a Chinese character sequence, based on the MPs of each Chinese character, a much smaller recognition net can be constructed. Take an example of a typical text sentence "為母說法", which is shown in Fig. 2. We call such a net as sausage net, which is named for its shape like a sausage. Higher recognition accuracy can be expected due to the smaller perplexity in the recognition net.

3.1. The Pronunciation Lexicons, Recognition Nets and Results

The **Formosa Lexicon** could be used for a wide range of applications and tends to have a higher number of multiple pronunciations in Taiwanese [2]. However, some pronunciations do not appear in the Formosa Lexicon due to pronunciation variations. Thus, the second lexicon, called the **Sutra Lexicon**, is derived from the Sutra itself to study what variations exist from the Formosa lexicon to Sutra Lexicon. The performance of phonetic transcription using the Sutra Lexicon is looked upon as the upper bound performance for the phonetic transcription. In addition to the above two separate lexicons, a combined lexicon is called the **Enhanced Lexicon** for convenience. In recognition nets, the first is the free-syllable net, denoted as the Free-Syl-Net. The other three search nets are the sausage nets were constructed by filling in each node of the net with the corresponding multiple pronunciations of each Chinese character from each of the three pronunciation lexicons. The nets are denoted the General-Sau-Net, Specific-Sau-Net, and Enhanced-Sau-Net for the Formosa Lexicon, Sutra Lexicon, and Enhanced Lexicon.

With the four search nets and acoustic models, the recognition results are shown in Fig. 3. In addition, we also show the result of only language, called G2P, with unigram. Through observing the experimental results, neither G2P with unigram nor Free-Syl-Net with adaptation model can reach acceptable performance. Therefore, it is necessary to integrate the language and acoustic parts. The General-Sau-Net could compete with the Specific-Sau-Net. Thus, if the speaker independent model could be adapted using some phonetically transcribed speech data, the adapted speaker independent model under the General-Sau-Net would be suitable for phonetic annotation task. Although some pronunciations of the Buddhist Sutra Chinese characters do not appear in the Formosa Lexicon, the performance of the Formosa Lexicon Sausage Net degrades not much more than the Sutra Lexicon Sausage Net. So far, the Enhanced Lexicon Sausage Net includes all possible pronunciations of Sutra Chinese characters, but it may increase the perplexity of the search net. Practically, some errors result from pronunciation variations by our speech data observation. Therefore, we determined that the performance would get better by trivial adaptation of the Formosa Lexicon

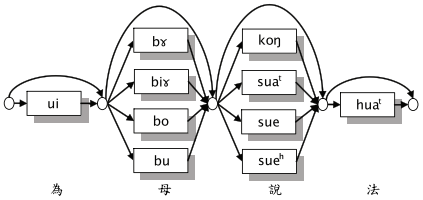

Fig. 2. The net is constructed from the multiple pronunciations of each Chinese character from our Formosa Lexicons.

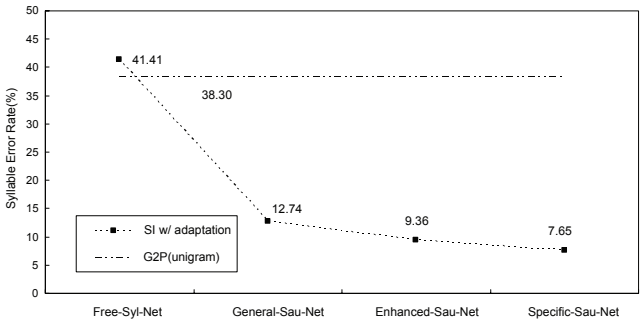

Fig. 3. Syllable error rate (SER) under four searching nets. See text in subsection 3.1 for notations.

Sausage net.

4. INCORPORATING PRONUNCIATION VARIATION RULES

Because insufficient coverage of pronunciations in the search net will severely degrade the recognition performance, some approaches to extend the pronunciation coverage will be considered to help the overall performance. The simple way to adopt the methodology of pronunciation variation is to expand the pronunciation lexicon using variation rules of the form $LBR \rightarrow LSR$ 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 [3]. 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 shown in Table 1.

A small portion of the ForSDAT-02 was then manually checked and the phonetic transcription of the transcript "corrected" according to actual speech. The triphone-level confusion table is built and used as a direct knowledge source to derive the PV rules, where each cell in the table is looked upon as a rule. The enormous number of rule set selections is 2^{P^2}, where P is the number of triphone. To make the problem more solvable, some specially designed algorithms should be developed. The mathematic definitions of the 3 kinds of stastical measures are as follows:

1. Joint probability (JP) 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 (CP) of the surface form pronunciation s_j, conditioned on the base form pronunciation b_i, $p(s_j|b_i) = n_{ij}/N_i$.

3. Mutual information (MI) of the base form pronunciation b_i, and the surface form pronunciation s_j, $I_{ij} = n_{ij}/N \cdot log(N \cdot n_{ij}/(\sum_i n_{ij} \cdot \sum_j n_{ij}))$.

where 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. We select those pairs (i, j) with higher scores of $p(b_i, s_j)$, $p(b_i, s_j)$ and I_{ij} to extend the sausage net pronunciation.

We adapted the Formosa pronunciation lexicon according to different pronunciation variation rule sets. The SER was **12.74%** before the application of the pronunciation variation rules as shown in Fig 3. This would be looked upon as the performance of the baseline setup in this section. In Fig. 4, we could observe that it is truly helpful to decrease the SER by increasing the search net coverage via the PV rules.

It is interesting to point out that, in Fig. 4, choosing different statistical measures will influence the achievable lowest SER and also the speed of decrease in SER. In these experiments, we found that MI is the best in terms of the rate of decrease in SER or the achievable lowest SER. In JP-based method, sometimes the insignificant and harmless PV-rules might get the higher conditional probability due to few base-form observations. The PV-rules for the CP-based method might not increase the perplexity but lead to the slowest convergence among the three methods. In the MI-based method, the formula could avoid slow convergence using the Joint-Probability as weight when the base-form would get few variations. Consequently, the error rate of the MI rank converges most quickly and the performance of the MI method in error reduction was also better than JP and CP methods, respectively.

Another interesting point was that the SER will possibly increase if too many PV rules were applied. For example, the lowest SER is achieved by applying a few rules when MI was adopted as the ranking measure. However, after applying more rules, the SER did increase! It will even become worse than that in the baseline experiments. This means that some "bad" pronunciation variation rules may lead to a performance reduction. Therefore, it is important to determine "good" rules and choose them such that the optimal performance could be achieved as soon as possible.

Extending the search 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 search net in Fig. 4. Regardless what methods we use, the differences in the perplexity values from the best results among the three methods were always slight. That means too many rules may lead to more real pronunciation coverage, but the performance may improve slightly or even decrease progressively.

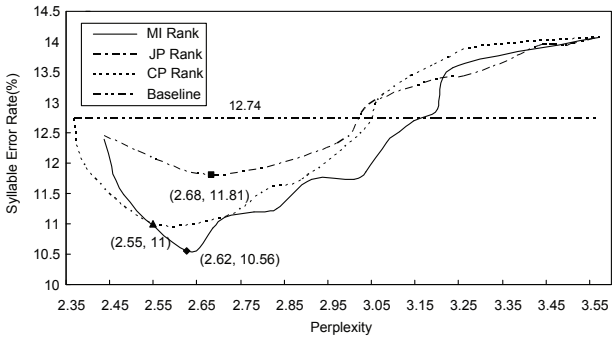

Fig. 4. The recognition result (syllable error rate) v.s. the perplexity sorted according to different measures, including MI, JP, and CP as well as the Baseline criterion

The perplexity is a good measure to evaluate the search net in obtaining the best results.

5. 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 MPs. By using a MP lexicon, a transcription error rate of 12.74% was achieved. In addition, the trivial adaptation of general purpose sausage net with pronunciation variation rules was used instead of global pronunciation lexicon modification. Further improvement of an error rate reduction of 17.11% could be achieved. Although the proposed technique was developed for Taiwanese speech, but 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.

6. REFERENCES

[1] Kim, D. Y., et al. Development of the CU-HTK 2004 Broadcast News Transcription Systems. ICASSP 2005, pp. 861-864.

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

[3] Saraclar, M., et al. Pronunciation change in conversation speech and its implications for automatic speech recognition. Computer Speech & Language 18, pp. 375-395, 2004.

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