Data-Driven Approach to Pronunciation Error Detection for Computer Assisted Language Teaching

Min-Siong Liang[1], Zien-Yong Hong[2], Ren-Yuan Lyu[2], Yuang-Chin Chiang[3]
[1.] *Dept. of Electrical Engineering, Chang Gung University, Taoyuan, Taiwan*
[2.] *Inst. of Computer Science and Information Engineering, Chang Gung University,Taiwan*
[3.] *Inst. of Statistics, National Tsing Hua University, Hsin-chu, Taiwan*
E-mail: {minsiong, renyuan.lyu}@gmail.com Tel: 886-3-2118800 ext 5967

Abstract

This paper describes an approach to pronunciation error detection for Computer-Assisted Pronunciation Teaching (CAPT). We focus on how to find the real pronunciation of the user. The data-driven based method was used to generate pronunciation errors hypotheses instead of knowledge-based method. In the experiment results, the error rate of pronunciation detection can achieve 10.56%. Finally, we applied this technique into our CAPT system.

1. Introduction

This paper describes an approach to pronunciation error detection for Computer-Assisted Pronunciation Teaching (CAPT). For CAPT, many researchers have developed it using speech recognition and synthesis techniques [1]. So far, most approaches for CAPT used the Hidden Markov Models (HMM) log-likelihood-based algorithm score [2], but few could detect and verify error for users. However, the report by Neri et al. [3] claimed that the effectiveness of learning depended on the corrective feedback of a CAPT system, which needed a precise pronunciation error detector. In addition, the pronunciation errors hypotheses often used linguistic knowledge, which was often language-dependent and derived by more than one linguists [4], but the linguistic knowledge was sometimes contradictory with each other. In this paper, therefore, we used data-driven based method instead of knowledge-based method for generation of pronunciation errors hypotheses and proposed a new framework for CAPT.

We were trying to incorporate our approach to our mother-tongue language, i.e. Taiwanese. Unfortunately, due to lack of elementary education for this language, new generation in Taiwan can not speak and listen to it. Although Taiwanese uses Chinese characters as a part of the written form, with its own phonetic system, it 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) [1]. 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. Finally, we also needed to find error patterns when students, who spoke Mandarin in daily life, started to learn Taiwanese. Therefore, it might be the better way to solve this problem using the data-driven approach.

In this paper, we focus on how to find the real pronunciation of the user and the detail will describe as the following sections.

2. The Pronunciation Error Detection using Speech Recognition Technique

The flow chart shown in Fig. 1 is the framework of phonetic transcription of utterances using the speech recognition technique. While the input is a speech waveform of a student and Chinese text of the lecture, the output is a phonetic transcription corresponding to pronunciation of the user. The entire framework can be divided into two major parts, i.e. an acoustic part and a language part.

Based on flow chart in Fig. 1, we define: $\underline{s}$ is the syllable sequence, while $\underline{c}$ and $\underline{o}$ are the input acoustic sequences and augmented character. The phonetic transcription target is to find the most probable syllable sequence $\underline{s}$* given $\underline{o}$ and $\underline{c}$. The formula is:

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

where $\underline{c} \in \underline{C} = \{\underline{c} \mid \underline{c} = c_1^N = c_1....c_N, c_i \in C\}$, c_i is an arbitrary Chinese character, C is the set of all Chinese characters. $\underline{s} \in \underline{S} = \{\underline{s} \mid \underline{s} = s_1^N = s_1....s_N, s_i \in S\}$, s_i is an arbitrary Taiwanese syllable, S is the set of all Taiwanese syllables. Use the Bayes theorem and assume that the acoustic sequence $\underline{o}$ is dependent only on the syllable sequence $\underline{s}$. Eq. 1 could be simplified as:

$$\underline{s}^* = \underset{\forall \underline{s} \in \underline{S}}{\arg\max}\, P(\underline{s} \mid \underline{c}) P(\underline{o} \mid \underline{s}) \quad (2)$$

The first term, $P(\underline{s} \mid \underline{c})$, of Eq. 2 is independent of $\underline{o}$ and plays the major role in the language part of the recognition scheme. The second term, $P(\underline{o} \mid \underline{s})$, is the probability of observation given the syllable sequence and plays the major role in the acoustic part.

For the acoustic part, we use Maximum Likelihood Linear Regression (MLLR) to adapt speaker independent models. The features were extracted into vectors of 48 dimensional MFCC plus 4 dimensional energy.

For the language part, it could be modeled as pronunciation to phonetic transcription problem. We consider the lexical pronunciation as the correct pronunciation. In addition, even the best pronunciation lexicon would miss the true pronunciation for a certain Chinese character. We also viewed the pronunciation variations as error patterns. To address this issue, the pronunciation variation rules, i.e. error patterns, would be incorporated in the searching net.

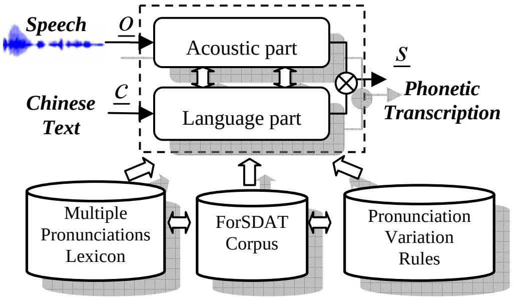

Fig. 1. The flow chart of the phonetic transcription of Taiwanese pronunciation error detection incorporating pronunciation variation rules

3. The Recognition Nets

3.1. The Correct Pronunciation Nets (CPN) with MP Lexicon

The **Formosa Lexicon** could be used for a wide range of applications and tends to have a higher number of multiple pronunciations in Taiwanese [5]. Given a Chinese character sequence based on the MPs of each Chinese character, a much small recognition net can be constructed. Take an example of a typical text sentence "爲母說法", which is shown in Fig. 2. We call such a net as sausage net, which is named for its shape like a sausage and the basic net could be considered as correct pronunciation net.

If the likelihood of utterances is high through the basic net, the pronunciation of student can be considered as standard pronunciation. However, we can not spot any errors when the likelihood is under the threshold. Therefore, it is necessary to incorporate the error patterns, i.e. pronunciation variation rules, to detect the real pronunciation of users.

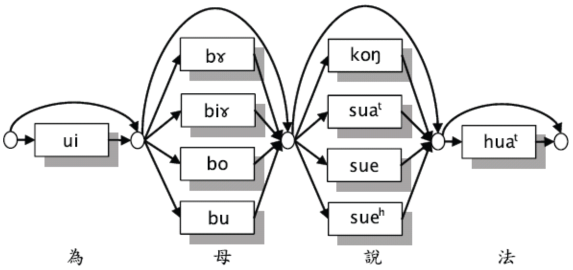

Fig. 2. The phonetic symbols used here are IPA followed by a digit representing one of several tone classes of the Taiwanese language.

3.2. Confusion Table Construction for Deriving PV 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". To derive such rules, a speech corpus with both canonical pronunciation and actual pronunciation is necessary. It could be shown that as long as the pronunciation rules could be well designed, the phonetic transcription performance would be effectively improved.

We choose a subset of ForSDAT [5], which contains 19731 utterances, to derive PV rules. The speech is recorded by first prompting a transcript to the speakers. Although the prompted transcript in Taiwanese text is shown with phonetic transcription, we do observe variations in the recorded speech. The speech data was then manually checked and the phonetic transcription of the transcript "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 1, which is called the tri-phone level confusion table.

	i-ng	*i-n*	…	s_j	…	*bh-o*	
bh-er	0	0	…	n_{1j}	…	30	267
…	…	…		…		…	
b_i	n_{i2}	n_{i3}	…	n_{ij}	…	$n_{i,p-2}$	N_i

…			…		…		
a-m	0	0	…	n_{Pj}	…	0	
	1315	1102		M_j		107	N

Table 1. Triphone-level confusion table, where the notations were described in section 3.2

3.3. Ranking PV Rules with Data-driven Method

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 b_i, and the surface form pronunciation s_j. The mathematic definitions of the above 3 measures are as follows:

(1) Joint probability of b_i and s_j,

$$p(b_i, s_j) = n_{ij} / N$$

(2)Conditional probability of b_i and s_j,

$$p(s_j \mid b_i) = n_{ij} / N_i$$

(3) Mutual information of b_i and s_j,

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

In all 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

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

$p(b_i, s_j)$ 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.

4. Evaluation of the Data-Driven methods

The testing data was the spoken Taiwanese corpus of Buddhist Sutra (written collections of Buddhist teachings), which is collected by a nun. There are 533 utterances in this speech data with total length of about 46 minutes. 502 utterances, which include 5909 syllables, are randomly chosen and reserved for testing while as another 31 utterances are used for acoustic model adaptation.

The experiment results were shown in Fig. 3. Under the speaker adaptation models, the result was 12.7% with the CPN. The adapted speaker independent model under the correct pronunciation net could be considered as the baseline of the pronunciation detection task.

It is interesting to point out that, in Fig. 3, choosing different statistical measures will influence the achievable lowest 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 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 performance of the MI method in error reduction was also better than JP and CP methods, respectively. Then, the error rate reduction of MI method was 17.11%.

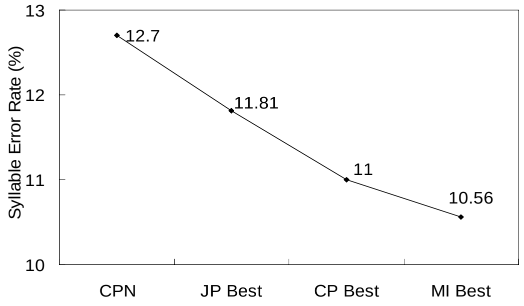

Fig. 3. The experiment results of CPN (baseline) and three best results using different data-driven methods.

5. Conclusion

We have proposed a new approach to address the pronunciation detection for learning Taiwanese pronunciation. By using a MP lexicon, a transcription error rate of 12.7% was achieved. In addition, the adaptation of correct pronunciation net (CPN) 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. Finally, we have also applied this technique to our CAPT system.

6. References

[1] M.-S. Liang, et al., “A Taiwanese Text-to-Speech System with Applications to Language Learning”, In Proc. ICALT 2004, Joensuu, Finland, (2004).

[2] S. Wei, et al., “Automatic Mandarin Pronunciation Scoring for Native Learners with Dialect Accent”, In Proc. Interspeech 2006, Pittsburgh, Pennsylvania, 2006.

[3] A. Neri, et al., “ASR-based Corrective Feedback on Pronunciation: does it really work?”, In Proc. Interspeech 2006, Pittsburgh, Pennsylvania, 2006.

[4] J.-C. Chen, et al., “Formant-Based English Vowel Assessment for Chinese in Taiwan”, In Proc. Interspeech 2006, Pittsburgh, Pennsylvania, 2006.

[5] R.-Y. Lyu, et al., “Toward Constructing A Multilingual Speech Corpus for Taiwanese (Minnan), Hakka, and Mandarin”, IJCLCLP, Vol. 9, No. 2, August 2004, pp. 1-12.