Pronunciation Error Detection for Computer Assisted Pronunciation Teaching in Mandarin

Min-Siong Liang[1], *Jian-Yung Hung*[2], *Ren-Yuan Lyu*[3], *Yuang-Chin Chiang*[4]

[1]Dept. of Electrical Engineering, Chang Gung University, Taoyuan, Taiwan
[2]Advanced Technology Center, Industrial Technology Research Institute, Hsinchu, Taiwan
[3]Inst. of Computer Science and Information Engineering, Taoyuan, Taiwan
[4]Inst. of Statistics, National Tsing Hua University, Hsinchu, Taiwan
{minsiong,tdy.hung,renyuan.lyu}@gmail.com

Abstract

In this paper, we provided a strategy of error detection of pronunciation and applied it to the computer-assisted pronunciation teaching(CAPT) , especially in Mandarin language learning. In our system, it can be divided into two parts: the sentence verification(SV) and syllable identification(SI). First was used to ban off-task sentences. We used the likelihood ratio test, which was computed between the maximum probability of a result under two different hypotheses, i.e. null hypothesis and alternative hypothesis models, to verify the deviation degree and decide whether the student pronunciation is off-task. In SV part, the experimental results was significant and had 91.0% rate of F-score. The second part was applied to recognize the content of speech read by the speaker. The recognition net was built as a sausage shape with pronunciation confusion table corresponding to confusion error patterns. Then, the system could find out the wrong pronounced syllable for the appropriate feedback to correct the pronunciation of the users. In the stage of SI, the best detection rate had a F-score rate of 77.2%.

Index Terms: computer assisted language teaching (CAPT), pronunciation error detection, sentence verification, syllable identification, Mandarin

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 [**?**, **?**]. So far, most approaches for CAPT used the Hidden Markov Models (HMM) log-likelihood-based algorithm score [**?**, **?**, **?**], but few could detect and verify error for users. However, the report by Neri et al. [**?**] 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 [**?**, **?**, **?**], but the linguistic knowledge was sometimes contradictory with each other. In this paper, therefore, we try to integrate data-driven based and knowledge-based methods for generation of pronunciation errors hypotheses to propose a new framework for CAPT in Mandarin.

As Fig. **??** was shown, our goal was to filter out the off-task sentences in the first stage and recognize the confusion or target pronunciations in the next stage. Therefore, the pronunciation error detection can be divided into two parts: sentence verification and syllable identification. In the sentence verification part, we were trying to ban off-task sentences with likelihood ratio test by comparing the correct pronunciation model and anti-pronunciation model. In order to give the corrective feedback with the right syllable or word in the next stage, we hope that the system could track considerable misreading, miscue, even the off-task sentences more accurately. On the other hand, the system was expected to recognize the speaker's pronunciation in phoneme or syllable level in the syllable identification part. The proposed solution to this problem is to expand the searching net with pronunciation variation (PV) rules. Even native speakers would have a pronunciation variation (PV) due to individual habits and accents. Collecting the PV patterns as many as possible might be a better way to find appropriate PV rules. However, how to rank those collected rules became another important issue. Therefore, we not only collect variation patterns but incorporated the linguistic knowledge and the ranking method to build the pronunciation searching network.

In Taiwan, Mandarin is one of three major languages (Mandarin, Taiwanese and Hakka) and is widely used as the native tongue. So far, for second language pronunciation learning, most systems were developed in English and few were built in Mandarin. Therefore, we were eager to build a language learning system in Mandarin and develop a prototype system shown as the Fig. **??** incorporating the proposed method. The interface consists of two major parts as follows:

One part includes prompt sheet block and assessment block. All learning materials are shown in this block. Each student can choose any sentence that they would like to learn, and then the system will pronounce the sentence. In addition, assessment block shows ratio of correct and error pronunciation that would help students realize the score in the specific lesson.

The other part is the pronunciations of the tutor and the student. The former demonstrates the native utterances for students to learn Mandarin. The user can listen to the pronunciation of arbitrary syllable shown in the interface. According to the pronunciation of the tutor, the student imitates the utterance, and then the system detect his/her pronunciation shown in the later block, so called "The pronunciation of the student".

2. The Sentence Verification for CAPT

To enhance error detection, the HMM-based sentence verification was used to filter out off-task pronunciations by likelihood ratio test. Given input speech to an HMM recognizer, let P_k be the most likely string of syllables by Viterbi decoding. P_k can

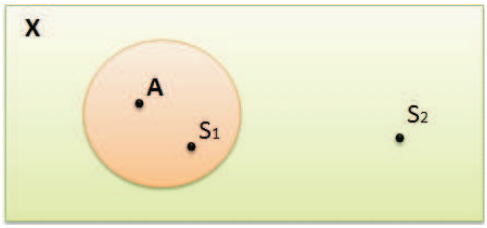

Figure 1: *The categories of pronunciation, where X, A, S_1 and S_2 can be represented as all possible pronunciations, target pronunciation, confusion pronunciation and off-task pronunciation respectively.*

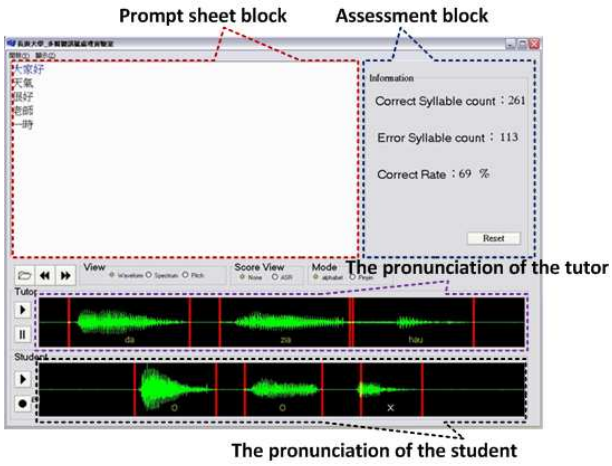

Figure 2: *The prototype interface of the proposed pronunciation learning system.*

be a concatenation of phoneme units which can be written as

$$P_k = p_1^{(k)} p_2^{(k)} ... p_{N_k}^{(k)} \quad (1)$$

where the phoneme string $p_1^{(k)} p_2^{(k)} ... p_{N_k}^{(k)}$ is the phoneme phoneme lexical representation of P_k, and N_k is the number of phoneme units comprising P_k. The log-likelihood ratio test can be written as follows:

$$log \frac{P(O|H_0(P_k))}{P(O|H_1(P_k))} = logP(O|H_0(P_k)) - P(O|H_1(P_k)) \quad (2)$$

where $P(O|H_0(P_k))$ and $P(O|H_1(P_k))$ were the likelihoods of the observation sequence O given the null hypothesis that P_k was spoken and the alternate hypothesis that P_k was not spoken respectively.

In order to find the better models for H_0 and H_1, we built four kinds of models, including Tri-phone models, garbage model (GM), pronunciation manner cluster(PCM) models and anti-PCM models. Table **??** showed how we clustered the Mandarin phonemes for PCM models. Instead, the anti-PCM models were trained by the whole observations except those of the phonemes of the same PCM cluster. Also, the GM model was trained by all observations of the whole phonemes.

2.1. Evaluation of the Sentence Verification

For the four kinds of acoustic models mentioned above, we chose speaker-independent HMM models trained by Formosa Speech Database (ForSDAT) [**?**]. The features were extracted into vectors of 48 dimensional MFCC plus 4 dimensional energy. For testing data collection, five sentences with 25 syllables were designed to cover all phonemes in Mandarin. Then, according to the table **??**, we replaced the five sentences with off-task(OT) or confusion(CF) syllables to form 160-sentence and 100-sentence prompt sheets for off-task and confusion pronunciation testing. All phonemes for each off-task syllable were chosen in the different clusters corresponding the original syllable. However, in order to form one confusion syllable, only one of phonemes of the original syllable would be replaced and chosen in the same cluster. For example, /bai/ can be the confusion pronunciation of /bʾai/ whereas the pronunciation of /xuaɴ/ was off-task . The distribution of these two kinds of prompt sheet was shown as Table **??**. Four native speakers were required to record the two prompt sheets and the statistics of the corpus was shown in Table **??**. Then, the criterion of rejection for a sentence occurred where a half of syllables was off-task. Finally, the experiment results were displayed in Fig. **??**. In addition, we also adopted F-measure to represent the performance of these acoustic models and found that the pair of tri-phone model versus anti-pmc was the best of them shown in Table **??**. Therefore, we decided the output of tri-phone vs. anti-PMC model as the input of testing data of the next stage, i.e. syllable identification.

Table 1: *The pronunciation manner cluster(PCM) transcribed in IPA in Mandarin.*

Consonant Class	Phone	Vowel Class	Phone
CS	b, bʾ, t, tʾ, k, kʾ	VI	i, y
CA	ts, tsʾ, tʂ, tʂʾ	VA	ɑ
CF	f, s, ʂ, z, x	VU	u
CN	m, n, ɴ	VO	o
CL	l	VER	ɤ,əˠ
		VE	ɛ

Table 2: *The distribution of off-task(OT) and confusion(CF) syllables in the two kinds of prompt sheets. Num. and sen. syl. are denoted as number, sentence and syllable. "Original:OT" means how many original and off-task syllables was in one sentence respectively.*

Original:OT	Num. of sen.	Original:CF	Num. of sen.
5:0	5	5:0	5
4:1	25	4:1	20
3:2	50	3:2	25
2:3	50	2:3	25
1:4	25	1:4	20
0:5	5	0:5	5

Table 3: *The statistics of the testing corpus. OT, CF, Sen., syl. and num. are denoted as off-task, confusion, sentence, syllable and number.*

	Num. of sen.	Num. of syl.	Time
OT corpus	640	3200	23 mins
CF corpus	400	2000	20 mins

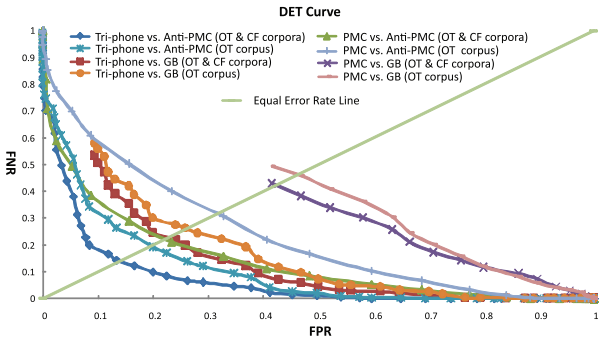

Figure 3: The experiment results of sentence verification under the OT and CF corpora with the combinations of Tri-phone, PMC, and GM, Anti-PMC models.

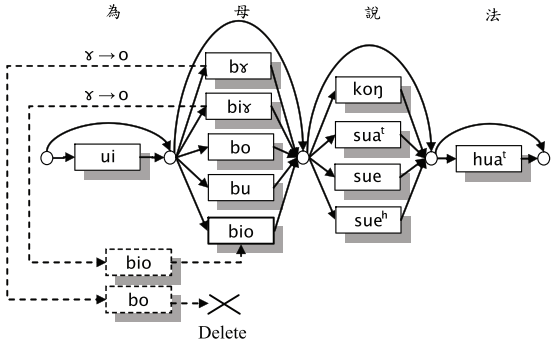

Figure 4: *An example of the extended sausage searching net. The net is constructed from the multiple pronunciations in lexicon and expanded using pronunciation-variation rules for each Chinese character according to the rule "/ɤ/→/o/".*

Table 4: *The best performance for the combination of PMC, Anti-PMC, GM and Tri-phone models under the OT and CF corpora.*

	Precision	Recall	F-Score
PMC vs Anti-PMC	81.0%	91.7%	86%
PMC vs GM	69%	100%	81.9%
Tri-phone vs Anti-PMC	88.9%	93.2%	91.0%
Tri-phone vs GM	81.1%	95.6%	87.8%

Table 5: *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 Syllables	179730	45865
Number of distinct triphones	1356	1194
Number of total triphones	555731	104894
Time(hr)	22.43	7.2

3. The Syllable Identification for CAPT

3.1. 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". Similar work for such an approach was shown in Mandarin [13]. To derive such rules, a speech corpus with both canonical pronunciation and actual pronunciation is necessary. An example is shown in Fig. **??**, 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 detection performance would be effectively improved.

We choose a subset of ForSDAT, called ForSDAT-02, to derive PV rules and the statistical information was summarized as in Table **??**. ForSDAT-02 is a speech database with rich biphone coverage. The speech is recorded by first prompting a transcript to the speakers. A small portion of 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 **??**, which is called the sentence-level confusion table. From the sentence-level confusion table, it is quite straight forward to construct another confusion tables in tri-phone level. This table is shown in Tables 3 as follows.

3.2. Ranking PV Rules with Data-driven based method and incorporating knowledge-based rules

First of all, the criterion should be adopted to choose the most significant rule sets. It is mutual information(MI) of the base form pronunciation and the surface form pronunciation. The mathematic definitions of the above measure is as follow:

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

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

In all the above equation, 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 addition, the variation rules also incorporated the linguistic knowledge[**?**]. The list of knowledge-based rules was shown in Table **??**. In Finally, in this stage, we also use F-measure (F-score) to evaluate our system under OT and CF corpora. The performance of syllable identification was presented in Table **??**.

4. Discussion and Conclusions

We have proposed a new approach to address pronunciation error detection for Computer-Assisted Pronunciation Teaching (CAPT). In our system, it can be divided into two parts: the sentence verification and syllable identification. To enhance error detection, we were trying four kinds of combinations to find the more better performance to filter out off-task pronunciations

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 (the base-form)	Manually corrected transcription (the surface-form)
e˩ bɤ˩ k'i˥˧ a˧	e˩ bo˩ k'i˥˧ a˧
gam˥ k'ai˩ ban˩ ts'en˥	gan˥ k'ai˩ ban˩ ts'en˥
siŋ˧ ua^{h}˩ hu^{h}˩ hua^{h}˧	sin˧ ua^{h}˩ hu^{h}˩ hua^{h}˧
⋮	⋮

Table 7: *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 $N = \sum_i \sum_j n_{ij}$.*

	b-ɤ	i-ŋ	i-n	...	s_j	...	b-o	
b-ɤ	237	0	0	…	n_{1i}	…	30	267
i-ŋ	0	1273	84	…	n_{2i}	…	0	1373
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
b_i	n_{i1}	n_{i2}	n_{i3}	…	n_{ij}	…	$n_{i,P}$	N_i
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
a-m	0	0	0	…	n_{Pj}	…	0	869
	241	1315	1102	…	M_j	…	107	N

by likelihood ratio test. In this stage, the experimental results under OT and CF corpora were significant and has 91.0% rate of F-score in combination of Tri-phone vs. Anti-PMC (anit-pronunciation manner cluster) models.

In pronunciation variation rules (PV-rules) selection, the data-driven variation rules, which were derived using mutual information(MI) measure, were used to extend more possible pronunciations. In addition to the data-driven measure, the linguistic knowledge was also integrated into the PV-rules. In the stage of syllable identification, the best detection rate had a F-score rate of 77.2%, which was the best among the other measures in this paper. Finally, we used these two components to build a prototype of CAPT as shown in Fig. **??**.

In the future, the more suitable acoustic models were still an unsolved problem in this research. The possible method might be to adapt the common acoustic models by the analysis of probability density function between mother-tongue and the second languages. In addition, the complex tone sandhi should also be accompanied with tone recognition.

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

5. References

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

[2] Abdou, S. M., et al.,"Computer Aided Pronunciation Learning System Using Speech Recognition Techniques", In Proc. Interspeech 2006, Pittsburgh, Pennsylvania, 2006.

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

[4] Witt, S. and S. Young,"Language Learning Based on Non-Native Speech Recognition", In Proc. Eurospeech 97, Rhodes, Greece, September 22-25, 1997.

[5] Franco, H., et al.,"Automatic Pronunciation Scoring for Language Instruction", In Proc. ICASSP 97, Munich, Germany, April 21-24, 1997.

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

[7] Tsubota, Y. et al.,"Recognition and Verification of English by Japanese Students for Computer-Assisted Language Learning System", In Proc. ICSLP 2002, Denver, USA, 2002.

[8] Tsurutani, C., et al.,"Development of a Program for Self Assessment of Japanese Pronunciation by English Learners", In Proc. Interspeech 2006, Pittsburgh, Pennsylvania, 2006.

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

[10] Kanokphara, S., et al.,"Pronunciation Variation Speech Recognition without Dictionary Modification on Sparse Database", In: Proc. ICASSP 2003, Hong Kong, 2003, pp. I-764-I-767.

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

[12] Lyu, Ren-Yuan, Min-Siong Liang, Dau-Cheng Lyu, and Yuan-Chin Chiang, "Advances in Chinese Spoken Language Processing", World Scientific, Chapter 17, pp. 387-406, 2007

Table 8: *The list for Knowledge-based rules including one-way and two-way patterns.*

One-way Pattern	Two-way Pattern
z̩→z	p↔p'
s̩→s	p↔t
ʦ→ts	p'↔t'
ʦ'→ts'	t↔t'
y→i	t'↔k'
	k↔k'
	n↔l

Table 9: *The best performance for Mutual-information-based and Knowledge-based measure under the OT and CF corpora.*

	Precision	Recall	F-Score
Mutual-Information	66.7%	91.5%	77.2%