USING EXTENDED LETTER-TO-SOUND RULES TO DETECT PRONUNCIATION ERRORS MADE BY CHINESE LEARNER OF ENGLISH

William Yang Wang, Lan Wang, Chongguo Li, Yajuan Huang

CAS-CUHK Shenzhen Institute of Advanced Integration Technology, Chinese Academy of Sciences, China
yw2347@columbia.edu {lan.wang, cg.li}@siat.ac.cn yj.huang@sub.siat.ac.cn

ABSTRACT

In this paper, we use extended letter-to-sound rules for automatic mispronunciation detection, aiming at checking pronunciation errors made by Chinese learners of English. The knowledge-based approach is used to generate extended pronunciation lexicon and incorporated into the HMM-based mispronunciation detection system. The pronunciation errors lead to misunderstanding of a word are expected to be identified. The TIMIT text prompts are used to collect data from Chinese university students, and the test set includes a total of 1900 sentences. Experiments show that the F-measure is about 0.86 at word level and about 0.91 at phone level. The system shows a high degree of accuracy in classifying correct and erroneous pronunciation.

Index Terms— Automatic Mispronunciation Detection (AMD), Extended Letter-to-sound Rules, Extended Pronunciation Lexicon

1. INTRODUCTION

For most Chinese learners of English, there are two main reasons to make pronunciation errors in their daily conversations. One is that the pronunciation of their native languages can directly influence the pronunciation of their second language. Another is that many non-native learners have imperfect understanding of letter-to-sound (LTS) rules and cannot correctly link graphemes of a word with its corresponding phonemes. Although learners can recognize the spelling of a word, they are not able to produce its correct pronunciation and especially, they have difficulties to pronounce confusable LTS patterns in the context of word.

Our previous study [1][2] indicated the effectiveness of using language transfer knowledge to find out the salient mispronunciations of Cantonese learners of English. However, the use of phonological comparisons for Cantonese learners of English may not be applicable to Mandarin speakers, since Chinese has at least seven main dialects [8] and phonological comparisons have great differences for these dialects. In this paper, we exploit letter-to-sound rules to improve mispronunciation detection for Chinese learners of English, which has commonly used in speech synthesis system [3][6]. Tejedor et al. [7] demonstrated the use of LTS rules in generating Spanish pronunciations. Hailemariam et al. [5] have also showed the efficiency of extracted LTS rules on Amharic, Hindi and Tamil language speech systems. Recently, the study [4] has applied LTS model to automatically generate lexical pronunciations.

In this paper, we start with the analysis of traditional English letter-to-sound rules [3], and then extend the pronunciation variations for the corresponding letter patterns in terms of manually detected pronunciation errors. Hereby, the TIMIT text prompts are used to collect the data of Chinese learners of English for training and testing. In this way, it is able to develop frequently made grapheme-to-phoneme confusions, and then generate extended LTS rules that summarize context-aware pronunciation errors. Based on the extended LTS rules, confusable phonemes of each word in model pronunciation lexicon are replaced by possible pronunciation variants and an extended pronunciation lexicon including predicted pronunciation variants of non-native speakers can be produced. The experiments were conducted on the test set containing 1900 sentences that recorded by 50 Chinese male and female learners of English. A promising result can be obtained by comparing the automatic detection to manual detection.

2. GENERATION OF EXTENDED LTS RULES

The proposed AMD system consists of an acoustic model trained on the TIMIT native-speaker corpus, the fixed-word sequence and an extended pronunciation lexicon. Based on automatic speech recognition, the phone-level transcription sequence is decoded for an input of non-native speaker's English. Then, through aligning and comparing the system phone-level sequence with model sequence generated via the pronunciation lexicon, it is able to recognize the erroneous pronunciations made by non-native learners. So, the procedures of generating extended pronunciation lexicon are vital since it predicts the possible pronunciation errors. In this paper, we focus on using LTS rules and investigate knowledge based approach to derive the extended pronunciation lexicon.

2.1. Training Set, Test Set and Annotation

The training set is part of Chinese Learners of English-Mandarin (CHLOE-Mandarin). The prompts of CHLOE-Mandarin corpus contain the texts used in TIMIT corpus.

First of all, a set of phonological rich sentences recorded by non-native speakers is selected as training set, the test set has no overlap with the training set in cases of speakers and sentences. The selected speakers are from major dialect regions of China. The training set is where we use LTS rules and compile the new linguistic knowledge of pronunciation variants from Chinese learners and then, produce extended LTS rules.

We invite a linguist to annotate all the non-native speech in both training and test sets manually, and keep a record of all the mispronounced phones and words, which is referred as Golden transcription [1]. Then, four American English native speakers were asked to double check the annotations.

2.2. Extended Letter-to-sound Rules

This knowledge-based approach requires a set of correct LTS rules, human annotations and model transcriptions, which contain correct pronunciations of each word.

A total of 329 LTS rules which was developed by [3] to translate English text into speech. We use these validated letter patterns to examine the training set and derive extended LTS rules from speech of non-native speakers. [3] offered a form of letter pattern to represent LTS rules. For instance, each rule may include a form like this:

#:[A]GE=/IH/

The left side of this form denotes a letter pattern when a letter A comes with one or more vowels and zero or more consonant ahead and followed by letter GE. If a word contains such pattern, letter A should be pronounced /IH/[1]. Special symbols to represent rules can be found in [3].

First of all, we use DTW and EM [9] to align graphemes and phonemes of each word in both human annotation and model pronunciation lexicon. It is to locate the letter pattern of the rule with its corresponding pronunciation. After alignment, each word entry in both human annotation and model lexicon have forms like this:

LAKE l ey k _

While the left part is a word and right part is the aligned pronunciation of this word. Here, "_" means silence. Then, each LTS rule in the LTS rule set is taken into aligned model lexicon to get their reference wordlist respectively. For example, as for the LTS rule ":[A]^+=/EY/", words satisfied this rule are bAke, lAke and so on.

If a word entry in human annotation appears in the reference wordlist with different pronunciation at the corresponding position of such letter pattern, it is then treated as a pronunciation variant of this word.

When we searched this pattern in human annotation, it turned out that many speakers pronounce the A as /AE/ in LAKE or /AE/ in BAKE. So, after evaluating the statistical data, an extended LTS rule, ":[A]^+=/AE/", is generated.

A pruning algorithm is developed which is based on the frequency of pronunciation variants in the human annotation. Since the generation process may lead to lots of implausible LTS rules. Therefore, extended LTS rules are generated, as shown in the Fig.1.

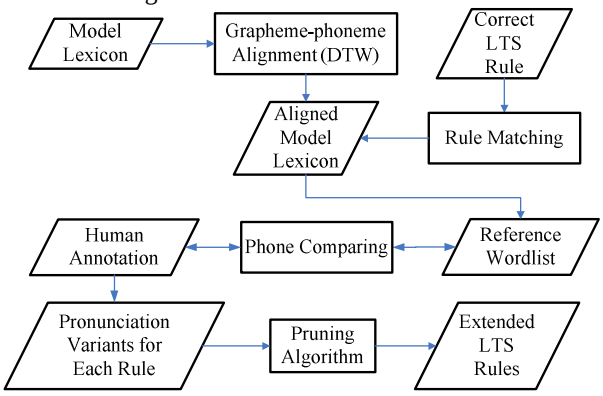

Fig.1. Extended LTS Rules Generation Framework

3. GENERATION OF EXTENDED PRONUNCIATION LEXICON

The extended pronunciation lexicon consists of correct pronunciation of each word in the model pronunciation lexicon, as well as predictable pronunciation variants of each word.

3.1. Extended Pronunciation Lexicon Generation

The generation of extended pronunciation lexicon requires a model lexicon, correct LTS rules and extended LTS rules. The entire process is analog to the generation of extended LTS rules. Except the correct pronunciation of target words in wordlist, those frequently made pronunciation errors are added into the dictionary in terms of the extended LTS rules. In particular, the rules with more restrictions are applied in prior to those with fewer restrictions. All redundant entries should be removed after generation.For example, after the extended pronunciation generation, the extended pronunciation lexicon has a form like this:

BAKE b ey k _
BAKE b ae k _
BAKE b ey k iy

3.2. ASR for Mispronunciation Detection

Using the extended pronunciation dictionary, an HMM-based mispronunciation detection [2] is performed to generate the phone sequences. The phone sequence alignment is conducted to compare the phone sequence recognition outcomes made by system, the model sequence generated from model pronunciation lexicon and the transcription of human annotation. A detailed alignment

[1]In this paper, we use Darpabet instead of IPA.

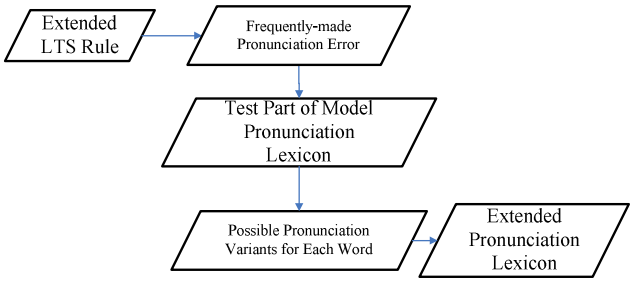

Fig.2. Extended Pronunciation Lexicon Generation

explanation of transcriptions can be found in our previous work [1].

4. EXPERIMENTS

In the experiments, it is expected to see whether or not the system can locate mispronounced words. Second, whether or not the system can diagnose the correct and erroneous phone within these words. Third, whether or not the system outcome can reach an agreement of pronunciation judgment with the human judge. Therefore, the performance evaluation of the system is conducted in both word level and phone level aspect. The acoustic models are the same as that used in [2].

4.1. Experiment Setup

The training set is a subset of 50 speakers (28 men and 22 women) selected from CHLOE-Mandarin corpus with the text prompts as those used in TIMIT training set. The test set contains the speech data from 50 speakers, where there is no speaker overlap in two sets. The Table 1 shows the word-level and sentence-level statistics of training and test set.

	Words			Sentences	
	Tokens	Unique	Error Tokens	Tokens	Unique
Training Subset	16734	2549	1389	1900	702
Test Subset	17321	2078	1289	1900	525

Table.1. The Statistics of Training and Test Subsets

Our original model pronunciation lexicon is the TIMIT dictionary. We split the TIMIT dictionary into training model lexicon and test model lexicon, according to the appearance of words in these two dataset. The number of words in test model pronunciation lexicon was 2078. After generating the extended LTS rules and applying to produce the extended pronunciation dictionary, 3153 to 8439 total entries were appeared in difference versions.

In Table 2, thresholds are the total number of this particular pronunciation variant happened in training set, the percentage refers to that stands in the entire pronunciation variants. The parameters are used for pruning the implausible pronunciation variations.

The evaluation of AMD system performance may be recognized as a classification problem, since the mis-

Threshold			Entries	
Version	Number	Percentage	Number	Ratio
V1	15	20%	8439	4.06
V2	20	35%	3799	1.83
V3	20	45%	3153	1.52

Table.2. The Scale of Extended Pronunciation Lexicon with Different Thresholds

pronunciation detection system works like a pronunciation classifier. So confusion matrix is introduced to analyze the system performance. A comprehensive explanation of confusion matrix used, including TA, FA, TR, FR, can also be found in our previous work [1].

4.2. Word-level Mispronunciation Detection Evaluation

First, the system performance is evaluated at word level and this means whether the system is able to detect word pronunciation precisely and accurately. Table 3 shows a classification of system outcome compared with model transcription and human judges.

	TA	FR	FA	TR	F0
V1	52.43%	27.39%	10.76%	9.42%	0.73
V2	70.53%	9.29%	16.43%	3.75%	0.85
V3	73.24%	6.57%	17.36%	2.82%	0.86

Table.3. The Word-level Classification Result

$$precision = \frac{TA}{TA + FA} \qquad accuracy = \frac{TA + TR}{R + A}$$

According to the word-level confusion matrix, the V1 word level precision is 82.98% and accuracy is 61.85%, while V2 are 81.11% and 74.28%, and V3 are 80.84% and 76.07%.

By comparing the human annotated transcription with model transcription, it is able to see the actual pronunciation performance of individual speakers. By comparing the outcome of AMD system and model transcription, the system performance is seen. By evaluating these two results, the agreement and difference between human judge and automatic system detection are exhibited.

The Fig. 3 indicates the word-level pronunciation errors located by human annotation. It is seen that the system can make judgment that is close to human's judgment. However, as for speaker 23, 30, 31 and speaker 33, system detections show differences with human annotations. We've checked the actual speech of these four speakers, it turned out that they sometimes pronounce odd phones that sounds like American English phones, but not entirely correct phones. For example, the speaker 23's pronunciation of the word "make" sounds like both /M EY K/ and /M AY K/. As for human annotation, the linguist put these pronunciations into correct category.

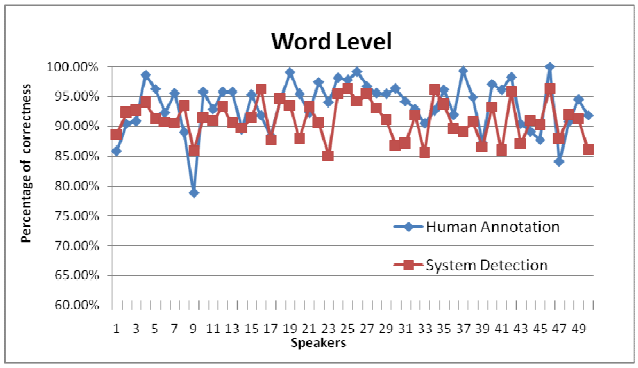

Fig.3. Word-level Performance Comparison for Individual Speakers: Human Annotation vs System Detection

4.3. Phone-level Mispronunciation Detection Evaluation

The phone-level evaluation is analogous to word-level evaluation and the difference is a smaller granularity is now considered.

	TA	FR	FA	TR	F0
V1	76.08%	8.99%	12.91%	2.02%	0.87
V2	80.79%	4.03%	14.19%	0.99%	0.90
V3	82.27%	2.37%	14.68%	0.67%	0.91

Table.4. The Phone-level Classification Result

According to the phone-level confusion matrix in Table 4, as for V1, the phone-level precision is 85.49% and accuracy is 78.10%, while V2 are 85.06% and 81.78%, and V3 are 84.86% and 82.94%.

As for Fig. 4, the trend of system detection and human annotation at phoneme level also coordinates. The most notable exceptions are speaker 8, 10, 12 and 42. After examining the recorded speech of these speakers, we found out that these speakers sometimes make unreasonable errors. For example, speaker 10 pronounced the word “vermouth” (/V ER M AX TH/) into “/V EH R IH M EH N SH AX N/”. This kind of utterance is unpredictable and should also be rejected in the first stage of computer-aided pronunciation training.

5. CONCLUSIONS AND FUTURE WORKS

This paper introduces novel approach of knowledge-based extended LTS rules in detecting mispronunciations. Among many factors influencing the performance of the AMD, this knowledge-based LTS rules approach using to generate extended pronunciation lexicon is crucial and effective. For further study, we still need to improve the training process to decrease the false rejection rate and increase the true acceptance rate.

6. ACKNOWLEDGEMENT

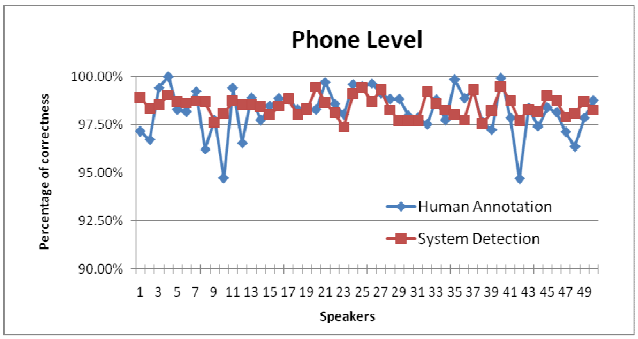

Fig.4. Phone-level Performance Comparison for Individual Speakers: Human Annotation vs System Detection

This work is supported by the National Nature Science Foundation of China (NSFC60772165), the Knowledge Innovation Program of the Chinese Academy of Sciences (KGCX2-YW-154) and the grant from the key laboratory of Robotics and Intelligent System, GuangDong Province (2009A060800016).

7. REFERENCES

[1]A.M. Harrison, W.Y. Lau, H. Meng and L. Wang, “Improving Mispronunciation Detection and Diagnosis of Learners' Speech with Context-Sensitive Phonological Rules based on Language Transfer”, In Proceedings of InterSpeech 2008, Page(s): 2787 – 2790, Brisbane, Australia, Sept. 2008

[2]L. Wang, X. Feng and H. Meng, “Automatic Generation and Pruning of Phonetic Mispronunciations to Support Computer-Aided Pronunciation Training”, InterSpeech 2008, Page(s): 1729 – 1732, Brisbane, Australia, Sept. 2008

[3] H. Elovitz, R. Johnson, A. McHugh, J. Shore, “Letter-to-sound rules for automatic translation of english text to phonetics”, IEEE Transactions on Acoustics, Speech and Signal Processing, Volume 24, Issue 6, Page(s): 446 – 459, Dec. 1976

[4]G. F. Choueiter, S. Seneff, J. R. Glass, “Automatic lexical pronunciations generation and update”, 2007 IEEE Workshop on Automatic Speech Recognition & Understanding, Page(s):225 – 230, Kyoto, Japan, Dec. 2007

[5]S. Hailemariam, K. Prahallad, “Extraction of Linguistic Information with the AID of Acoustic Data to Build Speech Systems”, 2007 IEEE International Conference on Acoustics, Speech and Signal Processing, Volume 4, Page(s): IV-717-IV-720, Honolulu, U.S.A., April 2007

[6]J. Lee and G. G. Lee, “A data-driven grapheme-to-phoneme conversion method using dynamic contextual converting rules for Korean TTS systems”, Computer Speech and Language, Volume 23, Issue 4, Page(s): 423-434, Oct. 2009.

[7]J. Tejedor, D. Wang, J. Frankel, S. King, J. Cola´s, “A comparison of grapheme and phoneme-based units for Spanish spoken term detection”, Speech Communication, Volume 50, Issue 11-12, Page(s): 980–991, Nov. 2008

[8]W. S-Y. Wang, “Languages and Dialects of China”, 1991

[9]A. W. Black, K. Lenzo, and V. Pagel. “Issues in building general letter to sound rules”, In Proceedings of the 3rd ESCA Workshop on Speech Synthesis, Page(s): 77–80, Jenolan Caves, Australia, Nov. 1998