MISPRONUNCIATION DETECTION BASED ON CROSS-LANGUAGE PHONOLOGICAL COMPARISONS

Lan Wang, Xin Feng and Helen M. Meng[1]

CAS/CUHK ShenZhen Institute of Advanced Integration Technologies, Chinese Academy of Sciences
[1]Human-Computer Communications Laboratory, The Chinese University of Hong Kong
lan.wang,xin.feng@siat.ac.cn,hmmeng@se.cuhk.edu.hk

ABSTRACT

This paper presents a method using speech recognition with linguistic constraints to detect the mispronunciations made by Cantonese learners of English. The predicted pronunciation errors have been derived from cross-language phonological comparisons, which are used to generate the erroneous pronunciation variations in a lexicon. The acoustic models are trained with native speakers' speech and used for recognizing the phone sequences, given the orthographic transcriptions. The experiments have examined that the agreement between automatic mispronunciation detection and human judges is over 84% for 21 Cantonese speakers.

Index Terms— mispronunciation detection, phonological comparison, speech recognition

1. INTRODUCTION

This work is investigating the automatic mispronunciation detection method to effectively highlight pronunciation errors made by Cantonese (L1) learners of American English (L2). The aim is to provide interactivity with regard to English teaching and self-learning in a computer-assisted language learning (CALL) system. In China, it would be especially useful to develop a CALL system with remedial instructions, since a large number of Chinese learners have no chance to talk with the native speakers to practice and then correct their pronunciations. However, most previous studies on CALL systems were designed to give a pronunciation measurement for non-native speakers [1, 7, 3, 9]. These methods incorporating with speech recognition techniques focus on accessing non-native speakers' pronunciation quality in a good or poor level. Other studies, such as [5], were conducted to discriminate a confusing pair of phonemes, e.g. the correct English pronunciation and its pronunciation marked by non-native speaker's accent.

In the paper, the target learners are adults who are native Cantonese and have learned English for some years. Automatic mispronunciation detection is then performed on the the continuous speech with various phonetic contexts. Since some English phonemes are missing from the Cantonese inventory, the Cantonese learners with accent often substitute for an English phoneme (that is missing from the Cantonese inventory) with a Cantonese one that has similar place or manner of articulation [6]. Such substitutions may lead to the confusion or misunderstanding of the English words. So the predicted pronunciation errors are derived from the cross-language phonological comparison, and then used to generate the erroneous pronunciation variations of each word in a constrained lexicon. Such extended pronunciation lexicon is then used with speech recognition system for automatic mispronunciation detection. The experiments are conducted with the English sentences recorded by 21 Cantonese speakers. The detection performance is then measured with the transcriptions for non-native database which have been annotated by human judges to reflect the actual spoken phone sequences.

The paper is organized as follows: Section 2 introduces the corpus and data collection, Section 3 describes the mispronunciation detection method. Section 4 is the experimental details and results. The conclusion and discussion are in Section 5.

2. CORPUS AND DATA COLLECTION

In this work, the acoustic models are trained with the TIMIT database, which contains the sentences covering a variety of phonetic contexts spoken by Americans in eight different districts. The Cantonese learners' recording were collected in a manner described in [6], which is referred to as CUHK collection. Three recording text prompts were used to collect speech from 21 Cantonese speakers and 5 native English speakers in the university. The recording texts include selected TIMIT sentences (284 sentences for each speaker) and the story of *The Northwind and The Sun*, which is commonly used by linguistics to exemplify languages.

Recordings of TIMIT sentences from the five native English speakers from CUHK were used to refine the acoustic models, and the speech of the story from Cantonese speakers was used to evaluate the mispronunciation detection performance.

3. MISPRONUNCIATION DETECTION

3.1. Methodology

Most pronunciation learning systems have used speech recognition techniques to access the pronunciation quality for language learners' word pairs or utterances. This paper is aiming to detect the possible pronunciation errors in the phoneme level, based on cross-language phonological comparisons and observations. We had compared the manner and place of articulations of vowels, diphthongs and consonants between Cantonese and American English phonetic inventory and identified the disparities across the language pair [6]. For the Cantonese learner with a strong accent, phonemes in the English phonemic inventory that are missing from the Cantonese inventory are often replaced by Cantonese phonemes in terms of production and perception. The substitutions always make confusion and misunderstanding of a English word.

To detect the mispronunciations made by non-native speakers, a phone-based recognition system built with English phone set is expected to recognize the pronunciation errors. However, the standard phone recognition systems obtain much higher phone error rate even for the native speaker. This makes it difficult to distinguish between pronunciation errors and recognition errors. In this work, the predicted mispronunciations are derived from phonological comparison between L1 and L2, which are used to add the pronunciation variations in a constrained lexicon. Mispronunciation detection is then performed by recognizing the most likely phone sequences using the trained acoustic models and an extended pronunciation dictionary with possible erroneous pronunciation variations, given the known word sequences.

3.2. Acoustic models

Using American English as the target language for most Cantonese learners, the acoustic models are built with TIMIT corpus recorded from native Americans. The HMM-based speech recognition techniques are used for pronunciation learning evaluation, where the cross-word triphone HMMs are trained with the reduced TIMIT phone set. The HLDA (Heteroskedastic LDA) [10] projection is performed to de-correlate feature vectors and reduce the feature dimensions.

The evaluation is performed on the speech of Cantonese, where the environment and device of data collection hereby are greatly different from the TIMIT database [6]. To reduce the mismatch of data collections between training and testing, the TIMIT data of native English speakers in the university is folded into the training set for acoustic modeling.

3.3. The extended pronunciation dictionary

The preliminary work in [6] has presented the comparative phonological analysis between L1 and L2, in a range of vowels, diphthongs and consonants. The salient mispronunciations made by Cantonese learner of English are then derived. In a summary, we tabulate the the possible phonetic confusions between L1 and L2 in a form of phone-to-phone mapping [1], which are then used to produce a lexicon with erroneous pronunciation variations.

Phone	Substitution	Deletion	Insertion
aa	ah		
ae	ah,aa, eh		
ah	ae		
ao	ow,aa,ah		
er	ah		
ih	iy		
iy	ih,ey		
uw	uh,ow		
uh	uw		
ow	ao,aa,ah	del	
aw	aa		
ay	iy, ih		
ey	ae, ih		
axr	ax,ah,ae,aa		
b	p		
d	t	del	d+ax
g	k	del	
k		del	
t		del	t+ax
p		del	
dh	d,t		
s		del	
sh	s		
th	t,d,f	del	
v	f,w		
z	s,t		
zh	s		
jh	g		
r	l,w	del	
y		del	
l	n	del	
n	l,ng		
ng	n	del	

Table 1. The confusable phone list

In Table 1, the confusable phone may be substituted, deleted or inserted in the continuous speech. For instance, the phone */ao/* may be probably substituted by the phone */ow/* when a Cantonese learner utters the word *north*. Some phones may be deleted in the continuous speech, which is marked by *del*. Some mappings only occur in a constrained case. For instance, the insertion */d/* $\rightarrow$ */d ax/* or */t/* $\rightarrow$ */t ax/* only happens when */d/* or */t/* is the final phoneme of a word.

Usually, only the correct pronunciations will be involved

[1] Henceforth we will use Darpabet instead of IPA

in a dictionary for speech recognition. To predict the possible mispronunciations in a word context, the above phone-to-phone mappings are used to generate an extended pronunciation dictionary. Each confusable phone is then mapped to zero (deletion), one (substitution) or more phones(insertions). For example, the word "there" */dh eh r/* may have confusion "dare" */d eh r/* with regard to the possible mapping between */dh/* $\rightarrow$ */d/*.

Applying the mappings for each phoneme in a word will produce a huge number of pronunciation variations in the dictionary, most of which are not real pronunciations of English words. So the pruning is required to remove those words that less probably occurred to enable the effective recognition.

3.4. The transcriptions and performance measures

To assess the pronunciation learning in a CALL system, the human judgment is required to score the pronunciation quality of non-native speakers. Pronunciation scoring mechanisms have been developed not only at the word-level, but also the phone level [7]. For mispronunciation detection proposed in this work, the phone-level pronunciation annotation is made by human judges for the evaluation. The human judges listen to the acoustic waveform with the phone transcriptions based on the right pronunciations, which is referred to as *target transcriptions*, and then locate pronunciation errors made by Cantonese learners. The manually annotated transcriptions are defined as *corrected transcriptions* [7], which give the phone sequences that L1 actually spoken.

In order to assess the effectiveness of mispronunciation detection, the performance measures compare transcriptions on a phone by phone basis. The phone-level similarity is computed between reference transcriptions manually made and the recognition outputs of the test speech. Two percentages are used as the standard speech recognition, the correctness refers to the percentage of all correctly detected phones, and the accuracy is calculated by taking account of the insertions.

4. THE EXPERIMENTS

This section presents experimental results for mispronunciation detection described in Section 3. All speech recognition is based on the cross-word triphone HMMs trained with the TIMIT corpus using the HTK toolkits.

The training set contains a total of 4620 sentences recorded by 462 native speakers from eight U.S. districts. In addition, the TIMIT speech of five native English speakers of CUHK collection has added to the training set, each speaker has 284 TIMIT sentences. The features were 13 cepstral coefficients and their derivatives, derived from MF-PLP analysis. The static cepstra were appended with 1st, 2nd and 3rd order derivatives to form a 52 dimensional features and then projected using a HLDA transform to 39 dimensions. A reduced TIMIT phone set containing 47 phones was used to construct the acoustic models. The state-clustered triphone HMMs were trained with 2000 distinct states and 12 Gaussian mixtures per state.

The test set contains the speech of *The Northwind and The Sun* recorded from 21 Cantonese learners. Each recording was processed and segmented into six utterances in continuous speech. For the test transcriptions of the story with a total of 70 words, the extended pronunciation dictionary has over 9000 pronunciations in all, using the phone-to-phone mappings described in Section 3.3. After pruning, unlikely pronunciation variations were removed from the extended pronunciation dictionary, so as to obtain a total of 349 pronunciations, including the correct ones.

The pairwise comparison has been made between the *target transcriptions* and *corrected transcriptions* of the testing speech.

	Correctness	Accuracy
target vs. corrected trans.	87.32%	86.57%

Table 2. Pairwise comparison between target transcriptions and and corrected transcriptions of the test speech

The figure indicates the pronunciation errors located by human judges, where about 13% substitutions have been made by Cantonese learners.

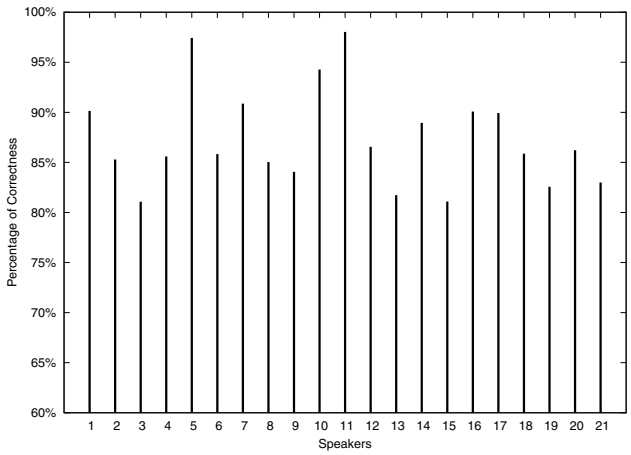

Fig. 1. The comparison between target transcriptions and and corrected transcriptions for each speaker

Some Cantonese learners pronounce fairly well, like speaker 5 and speaker 11, whose pronunciation substitutions are less than 3%. But one speaker has poor pronunciation with error rates at about 20%.

The trained acoustic models were then used with the extended pronunciation dictionary to run recognition, given the word sequences. Using the *corrected transcriptions* as the reference, the recognition outputs are compared with the references phone by phone. Both the fully extended pronuncia-

tion dictionary without any pruning (EPD v0) and the pruned dictionary (EPD v1) were used in the experiments.

	Correctness	Accuracy
EPD v0	79.80%	77.89%
EPD v1	84.47%	82.29%

Table 3. Pairwise comparisons between ASR outputs and corrected transcriptions

It is seen that the agreement between ASR outputs and human judgments is around 79-84%, The use of pruned dictionary did perform better than the use of full dictionary for recognition, since the redundant pronunciations cause underestimation of the posterior probability of the best pronunciation.

Since the human judgments are highly subjective, the strictness of human judges may lower the agreement between ASR and human judges. By analyzing mispronunciations pinpointed by human judges, it is also found that many pronunciation errors are out of the phone-to-phone mappings we derived from phonetics. Some mispronunciations made by Cantonese learners are due to the imperfect understanding of letter-to-sound rules. This can't be detected by the method proposed in this work, but the data-driven approach is under investigating to improve the mispronunciation detection.

5. CONCLUSIONS

This paper presents a method using speech recognition with linguistic constrains to detect the mispronunciations made by Cantonese learner of English. Unlike the previous studies on a CALL system, the phone-level pronunciation errors are proposed to be recognized, so as to provide instruction to language learners.

From the phonological comparisons of Cantonese versus English, the predicted phoneme pronunciation errors have been derived to generate an extended pronunciation dictionary for the recognition. The refined acoustic models were trained with native speakers' speech and used with the extended pronunciation dictionary for recognizing the phone sequences, given the orthographic transcriptions. The experimental results have shown the consistency of automatic mispronunciation detection and human judges. But there is still room to improve the mispronunciation detection method. Further investigations have been ongoing to refine the extended pronunciation dictionary with the data-driven approach.

6. ACKNOWLEDGMENTS

This work is supported by National Science Foundation of China (NSFC:60772165) and partly supported by CUHK Teaching Development Grant. The authors thank W.Y. Lau of CUHK to collect the data from Cantonese learners.

7. REFERENCES

[1] J. Mostow, A. G. Hauptmann, L. L. Chase, and S. Roth, "Towards a Reading Coach that Listens: Automated Detection of Oral Reading Errors", *Proceedings of the Eleventh National Conference on Artificial Intelligence (AAAI93), American Association for Artificial Intelligence*, Washington, DC, July 1993, pp. 392-397.

[2] W. Byrne, E. Knodt, S. Khudanpur, and J. Bernstein, "Is Automatic Speech Recognition Ready for Non-Native Speech? A Data Collection Effort and Initial Experiments in Modeling Conversational Hispanic English", *Proceedings of Conference on Speech Technology in Language Learning*, Marholmen, Sweden, 1998.

[3] B. Mak, M. H. Siu, M. Ng, Y. C. Tam, Y. C. Chan, K. W. Chan, K. Y. Leung, S. Ho, F. H. Chong, J. Wong, J. Lo "PLASER: Pronunciation Learning via Automatic Speech Recognition" *Proceedings of HLT-NAACL*, 2003.

[4] J. Mostow, "Is ASR Accurate Enough for Automated Reading Tutors, and How Can We Tell?", *Proceedings of International Conference on Spoken Language Processing ICSLP*, 2006.

[5] Y. Tsubota, T. Kawahara, and M.Dantsuji, "Recognition and verification of English by Japanese students for computer assisted language learning system", *Proceedings of ICSLP 2002,* pp.1205 1280, 2002.

[6] H. Meng, Y.Y. Lo, L. Wang and W.Y. Lau, "Deriving Salient Learners Mispronunciations From Cross-Language Phonological Comparison", *Proceedings of the ASRU 2007*, Kyoto, Japan, 2007.

[7] S.M. Witt and S. Young, "Performance Measures for Phone-level Pronunciation Teaching in CALL", *Proceedings of Speech Technology in Language Learning 1998* , pp.99-102, Sweden, 1998.

[8] Y.R. Oh, J.S. Yoon, H.K. Kim, "Acoustic Model Adaptation Based on Pronunciation Variability Analysis for Non-native Speech Recognition", Speech Communication, pp.59-70, vol. 49, 2007

[9] H. Franco, L. Neumeyer, Y. Kim and O. Ronen, "Automatic Pronunciation Scoring for Language Instruction", in *Proceedings of ICASSP 1997*, pp. 1471-1474, 1997

[10] N. Kumar, "Investigation of Silicon-auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition", Ph.D thesis, John Hopkins University, 1997