Automatic Generation and Pruning of Phonetic Mispronunciations to Support Computer-Aided Pronunciation Training

Lan Wang[1], *Xin Feng*[1], *Helen M. Meng*[2]

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

Abstract

This paper presents a mispronunciation detection system which uses automatic speech recognition to support computer-aided pronunciation training (CAPT). Our methodology extends a model pronunciation lexicon with possible phonetic mispronunciations that may appear in learners' speech. Generation of these pronunciation variants was previously achieved by means of phone-to-phone mapping rules derived from a cross-language phonological comparison between the primary language (L1, Cantonese) and secondary language (L2, American English). This rule-based generation process results in many implausible candidates of mispronunciation. We present a methodology that applies Viterbi decoding on learners' speech using an HMM-based recognizer and the fully extended pronunciation dictionary. Word boundaries are thus identified and all pronunciation variants are scored and ranked based on Viterbi scores. Pruning is applied to keep the N-best pronunciation variants which are deemed plausible candidates for mispronunciation detection. Experiments based on the speech recordings from 21 Cantonese learners of English shows that the agreement between automatic mispronunciation detection and human judges is over 86%.

Index Terms: mispronunciation detection, phonological comparison, speech recognition

1. Introduction

The aim of this work is to develop automatic instruments for language learning. We attempt to develop a mispronunciation detection system to effectively highlight pronunciation errors made by Cantonese (L1) learners of American English (L2). In previous research of computer-assisted language learning (CALL) systems, automatic assessment of pronunciation quality were widely studied to grade non-native speakers' pronunciations in a good or poor level. An automatic speech recognizer (ASR) was commonly used to test the pronunciation quality against the judgment of human raters, where the scores for either each phoneme or each word can give a rating of pronunciation [1, 8, 6, 2]. A recent study in [4] was conducted to discriminate a confusing pair of phonemes, e.g. the correct English pronunciation and its Japanese pronunciation marked by non-native speaker's accent. Another related work was presented in [7], where the monophone substitutions for English pronunciation variability from Korean speakers were obtained by analyzing phonetics and speech recognition results. The phoneme confusions were used to modify the statetying to adapt the acoustic models for accented speech recognition. In our research, a method is developed to identify the pronunciation errors at the phoneme level, so as to provide corrective feedback to support self-learning to improve English pronunciation using a CALL system.

The target learners are adults who are native Cantonese and have learned English for some years. It is observed that mispronunciations made in L2 are mainly due to the disparities at the phonetic and phonotactic levels across the language pairs. Since some English phonemes are missing from the Cantonese inventory, the Cantonese learners with accent often substitute for an English phoneme with a Cantonese one that has similar place or manner of articulation [5]. Such substitutions may lead to misunderstanding and confusion among the English words. For instance, the word *three* is commonly mispronounced as *free* by Cantonese (L1) speaker, where the missing phoneme */th/* is replaced by */f/*. This work focuses on automatically detecting such pronunciation errors caused by language transfer effects, using the speech recognition system with the predicted word mispronunciations on continuous speech.

We start with the cross-language phonological comparison, which summarizes missing English phonemes in the Cantonese inventory, in order to derive the *possible phonetic confusions* of Cantonese learners of English. Based on a set of phone-to-phone mapping rules, all confusable phonemes of a word in a lexicon are replaced by substitutions, deletions or insertions. In this way, the model dictionary is extended to include erroneous pronunciations to form the "extended pronunciation dictionary". In addition, the phone recognition results are analyzed to validate and complement the possible phonetic confusions. The extended pronunciation dictionary is then used with an ASR system trained with native speakers' speech to identify the possible mispronunciations. Since the combination of phone mappings for a multi-syllable word may produce many pronunciations that are implausible, we propose automatic pruning procedure based on the pronunciation confusion network to remove the implausible pronunciation variations in the extended pronunciation dictionary. Experiments are conducted with the English sentences recorded by 21 Cantonese speakers. The detection performance is then measured with respect to human transcriptions that reflect the actual spoken phone sequences.

The paper is organized as follows: Section 2 describes the system overview, Section 3 describes the procedures that generate the required dictionary to include the predicted mispronunciations, Section 4 presents the performance evaluation method, Section 5 provides experimental results. The conclusion and discussion are in Section 6.

2. The Mispronunciation Detection System

We investigate the use of automatic speech recognition (ASR) to detect the pronunciation errors made by Cantonese learners' of English. Detection is performed on continuously spoken English utterances. The target language is American English. Hence, we

develop an HMM-based speech recognition system using TIMIT corpus, which was recorded from native American English speakers.

A straightforward approach is to run phoneme-based recognition and then identify pronunciation errors in the learners' speech. Adaptation techniques can be used to reduce the mismatch between the training and testing sets to improve recognition accuracy, since the Cantonese learners' speech corpus was collected in a different setting compared to the TIMIT corpus. However, the phone recognition error rate is typically much higher than word error rate even for native speakers [10], which makes it difficult to distinguish between pronunciation errors and recognition errors. This means we can't directly apply ASR to mispronunciation detection.

Instead, we develop a mispronunciation detection system where the acoustic models and an extended pronunciation dictionary with possible erroneous pronunciation variations are used to recognize the most likely phone sequences, given the known word sequences. The key issue here is to predict the possible pronunciation errors and extend the standard TIMIT dictionary to include the erroneous pronunciations. Two methods have been investigated to derive the possible confusions at the phoneme level. One is based on the phonological comparisons between Cantonese learners of English. Another is a data-driven approach by performing automatic phone recognition on the learners' speech and analyzing the recognition errors to summarize phone-level confusions. The use of phonetic confusions can produce pronunciation variations for any word in a lexicon, so the mispronunciation detection system can be performed on the test data without any context constraints.

3. Generating The Extended Pronunciation Dictionary

3.1. Generation by phone-to-phone mapping rules

The study in [5] has compared the manner and place of articulations of vowels, diphthongs and consonants between Cantonese and American English phonetic inventory. The common mispronunciations are probably due to English phonemes that are *missing* in the Cantonese inventory. The *missing* vowels and consonants are observed to be replaced by the Cantonese vowels/consonants that are close in term of production and perception. For instance, the voiced plosives */b, d, g/* are present in English but not in Cantonese. Some Cantonese learners substitute with */p, t, k/* for these missing phonemes, which will result in mispronunciations of English words containing these voiced plosives. Possible phonetic confusions between L1 and L2 in the form of phone-to-phone mapping are derived from phonological comparisons. Accordingly, each confusable phone may be either substituted, deleted or inserted in the continuous speech. Most of mispronunciations are due to the phoneme substitutions, like the vowel confusion between */ao/→ /ow/*, which commonly occur when a Cantonese learner utters the word **north** [1]. The vowel insertion and consonant deletion may occur in Cantonese learners of English, due to the phonotactic constraints [5]. For instance, the consonant */t/* may be mispronounced by Cantonese learner with the deletion, which is marked by *del*. Some mappings only occur in a constrained case, like the insertion */d/ → /d ax/* or */t/ → /t ax/* only happens when */d/* or */t/* is the final phoneme of a word.

[1] Henceforth we will use Darpabet instead of IPA

3.2. Generation by data-driven technique

In addition to the phonetic confusions derived based on phonetics and phonology, the phone recognition errors of an ASR system can be used to get phone-level confusions. Phone-based recognition is run on the Cantonese learners' English, where testing adaptation is performed using MLLR transforms for each speaker. To derive the possible phonetic confusions, the recognized phone sequences are compared to the orthographic phone transcriptions. All the target phones with substitutions, insertions and deletions are concluded with the occurrence frequencies, which are computed as below:

$$S_{i/j} = \frac{Count\ of\ Substitution\ phone\ j}{Total\ Count\ of\ Phone\ i} \quad (1)$$

The substitution j can be "$-$" which means deletion, and "$+$" which means insertion. The minimum occurrence frequency is set to retain those most possible phonetic confusions. If a vowel is incorrectly recognized to be a consonant, or vice versa, the confusions are regarded as ASR recognition errors and then removed. The resulting phonetic confusions are then merged with those phone-to-phone mappings derived from phonological comparisons, to generate the predicted word mispronunciations.

3.3. Pruning the extended pronunciation dictionary

For speech recognition, only the correct pronunciations appear in the dictionary. To detect the mispronunciations made by Cantonese learners in a word context, the derived phone-to-phone mappings are used to extend the TIMIT dictionary to include predicted word pronunciation errors, which is referred to as the extended pronunciation dictionary. Each confusable phone of a word in a lexicon is 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/ → /d/*.

Since some confusable phones have more than two confusions, a large number of pronunciations may be produced after applying the phone-to-phone mappings to a multi-syllable word. Many of the variants in the extended pronunciation dictionary are implausible. Therefore, when Viterbi decoding is used with the fully extended pronunciation dictionary, Viterbi scores for the best phone sequences may be under-estimated. The recognition accuracy is thus degraded.

We develop a pruning procedure to remove the redundant pronunciations from the extended pronunciation dictionary. All pronunciation variations for each word are scored and ranked. The N-best pronunciations are then selected and the others are removed from the extended pronunciation dictionary. For this purpose, an implementation using the pronunciation confusion network is presented as illustrated in Figure 1.

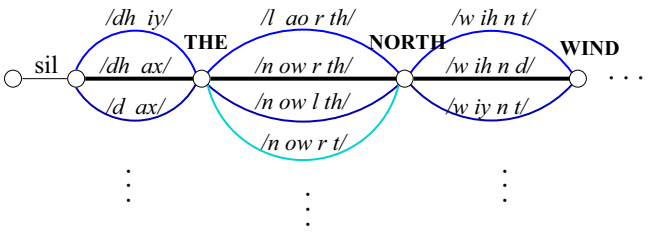

Figure 1: The pronunciation confusion network for pruning the redundant pronunciations.

Here, the node indicates the word with a fixed ending time and the links connected the nodes are all pronunciations for each node.

Viterbi decoding is first performed with the acoustic models and the fully extended pronunciation dictionary including all possible pronunciation variations, giving the word transcriptions. The starting/ending time of the best model sequences are then fixed. For each word, the Viterbi scores are calculated for each pronunciation variation in the particular staring/ending time, give the best model sequences of the other words. Thus, all the pronunciations of the same word in the continuous speech from multiple speakers are normalized and ranked. The pronunciation variants for each word in the extended pronunciation dictionary are retained if its pronunciation score is above the average pronunciation score, and the others are removed. After pruning, the fully extended pronunciation dictionary is reduced to the dictionary with most probably pronunciation variations.

4. 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 [6]. 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 corresponding to the right pronunciation of each word, 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* [6], which give the phone sequences that L1 actually spoken.

In order to evaluate 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: "Correctness" refers to the percentage of all correctly detected phones, and "Accuracy" is calculated by taking account of the insertions.

5. Experiments

This section presents experimental results for automatic mispronunciation detection using the extended pronunciation dictionary based on the possible phonetic confusions. The TIMIT database is used for acoustic model training and recognition. The Cantonese learners' recordings were collected in a manner described in [5], three recording text prompts were used including selected TIMIT test sentences and the story of *The Northwind and The Sun*.

5.1. Experimental set-up

The TIMIT training set contains a total of 4620 sentences recorded by 462 native speakers from eight U.S. districts. The acoustic 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 (Heteroskedastic LDA) [9] transform to 39 dimensions. A reduced TIMIT phone set containing 47 phones was used to build the cross-word triphone HMM-based recognition system. The state-clustered triphone HMMs were then trained with 2000 distinct states and 12 Gaussian mixtures per state.

5.2. Experimental results

The phone recognition was performed with the cross-word triphone HMMs and the phone pairs generated from TIMIT contexts, on the whole TIMIT test sets (native speakers). Meanwhile, recordings of TIMIT test utterances from Cantonese speakers were also decoded to derive the phonetic confusions, in addition to that from phonological comparisons. The recognition results of TIMIT test data of both native and Cantonese speakers are summarized in Table 1. Testing adaptation was used to estimate two MLLR transforms (one for silence, another for speech) per speaker, in order to improve the phone recognition accuracy of non-native speech.

	sub	del	ins	all
Native_spkrs	17.33%	9.03%	3.35%	29.71%
Cant_spkrs	32.10%	13.64%	9.65%	55.39%

Table 1: The phone error rates on both native and Cantonese TIMIT test data

It is seen that the phone error rate of the speech of Cantonese learners is much higher than that of native speakers, where the substitution of Cantonese data is nearly twice of that of native speech. Another reason is that TIMIT text prompts are unfamiliar for Cantonese learners to pronounce fluently and correctly. So the phone accuracy of the recognition system trained with native speakers' speech is low for the accented English.

For pronunciation evaluation, 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. Firstly, 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 and deletions have been made by Cantonese learners. The insertions are no more than 1% absolute.

Automatic mispronunciation detection was conducted by running forced-alignment with the acoustic models and the extended pronunciation dictionary based on possible phonetic confusions. The overall 51 phone-to-phone mapping rules derived from phonological comparisons were firstly used to generate the extended pronunciation dictionary (EPD v0). For the orthographic transcriptions with a total of 70 words, the fully extended pronunciation dictionary has over 6000 pronunciations in all. After pruning, unlikely pronunciation variations were removed from the extended pronunciation dictionary, and only 5% pronunciations were retained. Both the fully extended pronunciation dictionary and the pruned dictionary were used in the experiments. Using the *corrected transcriptions* as the reference, automatic mispronunciation detection outputs were compared to the references phone by phone. The agreement between automatic detection and human annotation is summarized in Table 3. Moreover, phone recognition results were also used to conclude 37 possible

phone-to-phone mappings, so as to generate the dictionary EPD v1. The merged phone-to-phone mappings were applied to produce the dictionary EPD v2, where the extra phonetic confusions from phone recognition results were added to that from phonological comparisons. The appropriate pruning was also performed respectively in the experiments, where the same settings were used to prune all the extended pronunciation dictionaries (EPD v1/v2/v3).

	size	Correctness	Accuracy
EPD v0	6674	80.57%	78.66%
pruned EPD v0	279	86.30%	83.05%
EPD v1	1038	82.50%	81.11%
pruned EPD v1	237	84.20%	82.17%
EPD v2	10772	77.90%	76.00%
pruned EPD v2	231	86.91%	83.93%

Table 3: Pairwise comparisons between ASR outputs and corrected transcriptions

In Table 3, it is seen that the disagreement between automatic mispronunciation detection outputs and human judgments is over 20% absolute, when using the fully extended pronunciation dictionary EPD v0. The pruning can effectively remove the redundant pronunciations from the EPD v0, the size of the EPD v0 is greatly reduced and the gain of agreement is 5.8% absolute.

The merged dictionary EPD v2 has an additional 3000 pronunciation variations compared to the EPD v0. An over large extended pronunciation dictionary may result in underestimation of the posterior probability of the best pronunciation, so the improvement can be observed only after pruning the EPD v2. About 115 pronunciation variations occur in both pruned EPD v0 and v2. The slight gain in both correctness and accuracy of the pairwise comparison between automatic detection and human annotation shows that the additional phonetic confusions from phone recognition results are helpful.

Since the human judges mark it as wrong if the phoneme can be perceived to be confusable with one of the common mispronunciations, the judgments might be strict. Different levels of strictness will be considered to improve the evaluation of mispronunciation detection. 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 (e.g. phonics). This case of mispronunciation is outside the scope of our current study.

6. Conclusions

This paper presents a mispronunciation detection system, using speech recognition with linguistic constrains to detect the phone-level pronunciation errors made by Cantonese learner of English. Unlike previous studies using ASR for CALL application, we focus on recognizing phone-level pronunciation errors, 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 acoustic models were trained with native speakers' speech and used with the extended pronunciation dictionary for recognizing the phone sequences, given the known word sequences. Moreover, the analysis of phone recognition results are conducted to obtain extra phonetic confusions for the predicted pronunciation errors. Experimental results have shown the consistency between automatic mispronunciation detection and detection by human judges. But there is still room to improve the mispronunciation detection method. Further investigations are ongoing to refine the acoustic models and improve the evaluation method for mispronunciation detection.

7. 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.

8. 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] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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

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

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

[10] L.F. Lamel and J.L. Gauvain, "High Performance Speaker-Independent Phone Recognition Using CDHMM", in *Proceedings of EUROSPEECH 1993*