ISCA Archive
http://www.isca-speech.org/archive

Speech and Language Technology in Education (SLaTE2009)
Wroxall Abbey Estate, Warwickshire, UK
September 3-5, 2009

Implementation of an Extended Recognition Network for Mispronunciation Detection and Diagnosis in Computer-Assisted Pronunciation Training

Alissa M. HARRISON, Wai-kit LO, Xiao-jun QIAN, Helen MENG

The Chinese University of Hong Kong, Shatin, N.T., Hong Kong
{alissa, wklo, xjqian, hmmeng}@se.cuhk.edu.hk

Abstract

This paper presents recent extensions to our ongoing effort in developing speech recognition for automatic mispronunciation detection and diagnosis in the interlanguage of Chinese learners of English. We have developed a set of context-sensitive phonological rules based on cross-language (Cantonese versus English) analysis which has also been validated against common mispronunciations observed from the learners interlanguage. These rules are represented as finite state transducers which can generate an extended recognition network (ERN) based on arbitrary canonical pronunciations. The ERN includes not only standard English pronunciations but also common mispronunciations of learners. Recognition with the ERN enables the speech recognizer to phonetically transcribe the learner's input speech. This transcription can be compared with the canonical pronunciations to identify the location(s) and type(s) of phonetic differences, thus facilitating mispronunciation detection and diagnoses. We have developed a prototype implementation known as the CHELSEA system and have validated the approach based on a new, annotated test set of 600 utterances recorded from 100 Cantonese learners of English. The approach achieves a false rejection rate (i.e. system identifies a phone as incorrect when it is actually correctly pronounced) of 13.6%; as well as a false acceptance rate (i.e. system identifies a phone as correct when it is actually mispronounced) of 44.7%. Among the detected errors, the system can correctly diagnose 54.8% of the mispronunciations.

1. Introduction

With the growing population of second language learners, there is a strong need for additional language learning resources. Kachru [1] estimates there are 533 million English learners in India and China alone—a number greater than the total population of the USA, UK, and Canada combined. With such a huge demand, there is an acute shortage of qualified teachers. Computer-assisted language learning (CALL) applications can supplement existing learning resources and provide unique benefits to learner in terms of accessibility, reduced anxiety, and individualized instruction.

Effective language learning tools, and particularly pronunciation training, needs to provide learners with detailed corrective feedback. Previous work has shown that automatic pronunciation scores at the word-level or sentence-level correlate highly with human raters but fail to lead to measurable improvement in learner's overall pronunciation [2]. However, locating mispronunciations at the phone-level to learners has been shown to lead to statistically significant improvement for the production of those targeted phones [3]. Moreover, diagnostic feedback to learners (e.g. "you inserted a vowel at the end of the word") has also been shown to lead to significant improvements in pronunciation training [4].

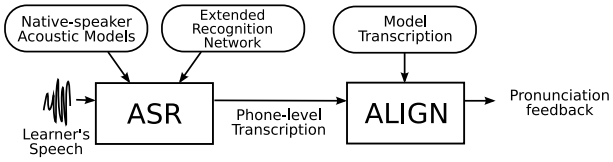

Figure 1: *Components of proposed computer-assisted pronunciation training tool*

Speech recognition systems must be specially designed for computer-assisted pronunciation training (CAPT) in order to support detailed corrective feedback while still obtaining satisfactory performance [5]. Although large vocabulary continuous speech recognition (LVCSR) systems are widely available in the commercial market, they are not necessarily appropriate for pronunciation training for non-native speakers. Generally, LVCSR systems are designed to *accommodate* a wide variety of accents and non-standard pronunciations. They are not intended to be used as a tool for discriminating phonetically similar pronunciations of a given word. Free phone recognition, in principle, could support a CAPT tool in providing detailed phone-level feedback to a learner. However highly accurate free phone recognition is still not possible for native-speech, and is only expected to be even more difficult for the interlanguage of second language learners. To address these problems, we attempt to develop a speech recognition system to support CAPT where the recognition network explicitly models common mispronunciations of the learner community. This approach allows us to develop a CAPT system which can detect and diagnose mispronunciations at the phone-level.

In this paper, we first give a brief overview of our proposed CAPT system design. Next we explain how the speech recognizer is built, with particular emphasis on the development of its *extended recognition network*. This is followed by an explanation of how the recognition output is processed to give feedback to the learner. The system is then validated with actual learner data from the CU-CHLOE corpus, in comparison with a free phone recognizer and a recognizer with a fully-informed recognition network. The paper concludes with our directions for future work.

2. Overview of CAPT system

The system flow of our proposed CAPT tool is as follows: (1) system prompts learner to speak a given utterance (2) learner records their speech (3) signal is recognized in according to the *extended recognition network* (4) recognized transcription

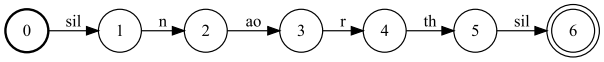

Figure 2: *Standard recognition network of "north"*

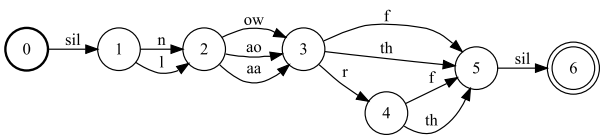

Figure 3: *Extended recognition network of "north"*

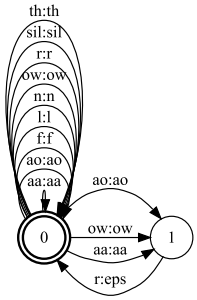

Figure 4: *Finite state transducer expressing r-deletion*

of learner is aligned with transcription of model native speaker (5) differences are highlighted as mispronunciations. An illustration of the system is provided in Figure 1.

For example, the system prompts the learner to speak an utterance that contains the word "north". The system uses a standard English pronunciation lexicon to build a standard recognition network of the model pronunciation(s) (Figure 2). This network is extended with common mispronunciations by composition with a finite state transducer [6] that models phonological processes in the learner community. This *extended recognition network* (Figure 3) is used by the speech recognition system in conjunction with acoustic models to output a phone-level transcription of the learner's speech, e.g. $/l\ ow\ f/$. This recognized transcription is aligned via a dynamic programming string alignment algorithm to find the differences between the learner's pronunciation and the standard pronunciation $/n\ ao\ r\ th/$. The CAPT system utilizes this information to provide corrective feedback to the user. For example, the system may inform the learner that they have failed to produce the $/r/$ and inform the learner to curl the tongue towards the roof of the mouth. This feedback may be further supplemented with visuals to illustrate the articulatory motions.

3. Automatic speech recognizer for CAPT

3.1. Extended Recognition Network

To characterize the phonological processes in Cantonese-speaking learners of English, we have gathered phonological rules from previous second language acquisition literature [7] and speech of 21 Cantonese-speaking learners of English reading "The North Wind and the Sun" (pilot collection of CU-CHLOE corpus) [8]. Commonly confused phones identified by [8] were further reviewed for contextual constraints and first language phonotactic constraints as discussed in [9]. From these analyses, a total of 51 phonological rules have been developed, in the following form:

$$\phi \rightarrow \psi \ / \ \lambda \_\_ \rho \tag{1}$$

This rule is read as follows: ϕ in the target language may be pronounced as ψ by the learner when following λ and preceding ρ. In our formulation of the rules, the context (λ and ρ) can include multiple phones or no phones and the following special symbols: # to indicate word boundaries, C to indicate any consonant, V to indicate any vowel, and F to indicate any fricative. On the other hand, ϕ and ψ in the rewrite mapping are restricted to a single phone. Phonological processes like deletion and insertion can still be indicated using the reserved symbol eps (e.g. $\phi \rightarrow eps$ indicates phone deletion). All phonological rules are defined as optional since learner speech is variable and they may not always mispronounce words. This means that a rule will generate at least two corresponding output forms for a single input form. The rules below model several common processes observed in Cantonese-speaking learners: substitute $/ow/$ for $/ao/$, substitute $/aa/$ for $/ao/$, delete $/r/$ following a vowel, substitute $/f/$ for $/th/$, and substitute $/n/$ for $/l/$ word-initially.

$$ao \rightarrow ow \tag{2}$$

$$ao \rightarrow aa \tag{3}$$

$$r \rightarrow / \ V \_\_ \tag{4}$$

$$th \rightarrow f \tag{5}$$

$$n \rightarrow l / \ \# \_\_ \tag{6}$$

In order to automatically extend the pronunciation network using these phonological rules, we represent them as finite state transducers (FSTs) using the open-source toolkit OpenFST [10]. An FST can be visualized as a directed graph with labeled transitions between states in the form $\alpha : \beta$ where α represents an input symbol and β an output symbol. It has an initial starting state (denoted by thick border) and accept states (denoted by a double-line border). If the input to the FST matches α on a transition from its current state, the machine moves to the next corresponding state and outputs β. If the machine reaches the end of the input while in an non-accept state, the output is blocked. A full description of FSTs and how they can represent phonological rules can be found in [6].

In this work, we adopt a simple albeit limited method for expressing phonological rules as an FST. The left-hand context of a rule is a series of identity mappings from the initial state to state n where n is the number of consecutive phones represented by λ. The rewrite mapping is a single transition from state n to $n + 1$. The right-hand context is again a series of identity mappings, where the last transition returns to the initial state. The optionality of the rules can be expressed by including self-transitions with identity mappings for every possible phone on the initial state. An example of an FST for $/r/$-deletion (Rule 4) is shown in Figure 4 where the possible phones are $\{ao, aa, f, l, n, ow, r, th, sil\}$ and $V = \{ao, aa, ow\}$.

It is important to note that expressing rules as an FST using this method is limited with respect to successive application of the same rule. For example in a hypothetical string $/ow\ r\ r/$, only the first $/r/$ can be deleted. While phonological processes in theory do not have this limitation (e.g. multiple successive consonant deletion), we find that our simple FST expression is still sufficient to represent the majority of mispronunciations in Cantonese-speaking learners (see Section 5).

Once each rule has been expressed as an FST, we can combine all rules into a single machine through a series of compositions (in the order of the rule listing). For example, if given three phonological rules R_1, R_2, and R_3, then the final transducer T modelling all processes is created as follows:

$$T = (T_1 \circ T_2) \circ T_3, \tag{7}$$

where T_1, T_2, and T_3 are the corresponding transducers of the rules. After composition, we have a single FST that models an ordered application of the 51 phonological processes we have identified. This FST is then composed with a standard recognition network to generate the ERN[1]. The standard recognition network previously given in Figure 2 and a FST representing Rules 2-6 will generate the ERN shown Figure 3.

3.2. Acoustic models

Acoustic models for the speech recognizer are cross-word triphone HMMs trained using the TIMIT TRAIN subset. The TIMIT corpus has been chosen because it is phonetically-balanced and has been hand-transcribed at the phonetic level. Its training data includes 462 American English speakers from 8 major dialect region each speaking 10 prompts (out of a total of 1718 distinct texts). Each of the HMMs are tri-state models with 12 Gaussian mixtures. Thirteen-dimension PLP features are used with first, second-order derivatives and cepstral mean normalization. Altogether there are 944 unique HMM states and 3338 unique models after state tying using a decision tree with phonetic context questions. The entire speech recognition system (training and testing) is implemented using the HTK toolkit from Cambridge University.

4. Pronunciation feedback

The recognition output of the speech recognizer is a phone sequence. To provide pronunciation feedback to the learner, this recognized phone sequence is automatically aligned with a model native-speaker pronunciation using dynamic programming. Phones in the learner pronunciation which differ from the model pronunciation can then be identified to the user and followed up with instructions for pronunciation improvement.

The system aligns the model pronunciation and the learner pronunciation utilizing phonetic features. We call this method a "phonetically-sensitive string alignment." This method is different from a standard string alignment algorithm which assigns a constant cost for each type of edit—insertion, deletion, and substitution—and returns the alignment with the minimal edit distance. Instead of a constant cost for substitution, our phonetically-sensitive alignment calculates a substitution cost based on the number of mismatched phonetic features similar to [11]. In our particular implementation we use 20 binary phonetic features adapted from [12]. Substitution cost is the sum of mismatched phonetic features and the insertion/deletion cost. The insertion and deletion costs are fixed to be approximately a third of the maximum substitution cost (7).

A review of selected alignments on a test set of 21 speakers by a linguist concluded the phonetically-sensitive alignment were more reflective of actual phonological processes than those alignments from a standard string alignment. An example of such an improvement is illustrated in Table 1. In the standard string alignment, the mispronounced vowel $/ow/$ is aligned with $/r/$ but the phonetically-sensitive alignment instead aligns $/ow/$ with $/ao/$. The first alignment implies the learner must insert a new vowel and change the second vowel to a rhotic, which is complex set of articulatory motions. But the second alignment enables us to simply instruct the learner to make two articulatory changes: lower their tongue to produce a lower vowel and then curl the tongue to produce a rhotic. We find that the phonetically-sensitive alignments are better motivated by linguistic theory and enables us to provide better feedback to the learner.

Table 1: *Comparison of standard string alignment for the word 'north' with constant substitution cost and phonetically-sensitive alignment*

	n	ao	r	th
Standard	l		ow	f
Phonetically-sensitive	l	ow		f

5. Validating the system

5.1. Methodology

The testing data includes 100 Cantonese-speaking learners of English reading "The North Wind and the Sun" from the CU-CHLOE corpus (disjoint with the previously mentioned pilot set). The passage is segmented into 6 sentences, providing a total of 600 utterances in the test set. The test set has been annotated at the phone-level by a phonetician. The annotations use a broad transcription method with the ARPABET phonemic symbols.

We compare the recognition accuracy of using our proposed method for building an *extended recognition network* with two other recognition networks: (1) *naive network* (2) *fully-informed network*. The naive network is a single state with a self-transition for each phone in the recognizer, i.e. equivalent to a free phone recognizer. The fully-informed network is built from the manually-annotated transcriptions of the testing data (i.e. the 100 Cantonese-speaking learners of English reading "The North Wind and the Sun"). It is called fully-informed as it includes all the pronunciations of the learners in the test set.

5.2. Results

Before comparing the recognition performance with the three types of recognition networks, we examined how well our extended recognition network modeled word mispronunciations in the test set. Of all the word tokens in the test set (n=11,285), 72.9% were modeled by the extended recognition network. Nearly two-thirds of the word tokens in the test set were mispronounced. When looking specifically at mispronounced words (n=7343), we found that 58.4% were included in the ERN. In contrast, a standard recognition network cannot model learner's speech as it only covers 34.9% of the learners word pronunciations, i.e. those correctly pronounced.

We then used the recognition networks to recognize the test set and measured the recognition accuracy by comparing the recognized transcription to the human annotation of the learners' actual speech. Unsurprisingly, the naive recognition network—equivalent to a free phone recognizer—had a poor performance of only 38.75% accuracy. The extended recognition network had 73.02% accuracy and the fully-informed

[1] When using OpenFST, the resulting ERN is encoded as a finite state acceptor in the AT&T FSM file format. This can be mapped directly to the HTK Standard Lattice File for use in the speech recognizer when the acceptor is deterministic.

network had slightly lower performance of 70.12%. The free phone recognizer performance is clearly unsatisfactory for a CAPT system, as the great majority of feedback to the learner would be incorrect. Thus, its mispronunciation detection and diagnosis performance is not further analyzed.

In Table 2, we compare the automatic mispronunciation detection and diagnosis performance of the extended and fully-informed networks. The performance is summarized as the false rejection rate (FRR), false acceptance rate (FAR), and diagnostic accuracy (DA). FRR measures the percentage of correctly pronounced phones erroneously rejected as mispronounced, while FAR measures the percentage of mispronounced phones erroneously accepted as correct. The diagnostic accuracy is the percentage of *detected* mispronounced phones that were correctly recognized (i.e. identical to human annotation).

Using the extended recognition network, the FRR and FAR are 13.55% and 44.72% respectively. Of the phones detected as mispronounced by the system, 54.80% of them were diagnosed accurately. The fully-informed network on the other hand has a FRR and FAR of 20.99% and 23.09%, respectively. Its diagnostic accuracy is 48.69%. Since the fully-informed network better reflects the pronunciations in the test set it has a lower FAR than the ERN. But the fully-informed performance is also vulnerable to having a overly bushy network due to idiosyncratic pronunciations, and thus its FRR and DA is actually worse than ERN.

Table 2: *Comparison of recognition performance using between the extended and fully-informed recognition networks*

Network	FRR	FAR	DA
Extended	13.55%	44.72%	54.80%
Fully-informed	20.99%	23.09%	48.69%

While we aim to minimize both error rates, FRR and FAR, there is an inherent trade-off between the two. We attach greater importance to FRR for the purpose of a CAPT since it is critical to avoid discouraging learners by rejecting correct pronunciations. On the other hand, failing to detect some mispronunciations does not have as severe consequences for the learner.

The performance of the ERN demonstrates that our system design can detect mispronounced phones with reasonable precision and even diagnose slightly more than half of the detected mispronunciations correctly. However, we acknowledge that there is much room for improvement in the system's performance. We see several areas where further gains could be made: utilization of posterior probability scores for classification decision [13], weighted recognition networks to reflect the relative probabilities of mispronunciations, and better acoustic models for non-native speakers via discriminative training and adaptation.

6. Conclusions and Future Work

We have proposed a system design for automatic detection and diagnosis of second language learners mispronunciations. Our approach represents common phonological processes in learners as finite state transducers, which can be used in turn to extend a standard recognition network with learners mispronunciations. This extended recognition network in conjunction with acoustic models trained on native speakers allows us to output a phonetic transcription of learners' speech. This recognized transcription is aligned with a model native speaker transcription to identify the location and type of mispronunciations to the learner at the phone-level. Furthermore, our approach has been validated on a large set of actual learner data. We are able to model the majority (75%) of learner's pronunciations from a set of 51 context-sensitive rules, while still obtaining recognition performance similar to a network which covers all learner pronunciations. In future work, we plan on testing the generalizability of our context-sensitive rules to data in other domains and investigating methods for automatic induction of an extended recognition network.

6.1. Acknowledgments

The authors would like to acknowledge Patrick Chu for his transcription work and review of phonetic alignments. We would also like to thank Dr. Pauline Lee of the CUHK Independent Learning Centre for her insights on common mispronunciations of Cantonese learners of English. This work is affiliated with the CUHK MoE-Microsoft Key Laboratory of Human-centric Computing and Interface Technologies. It is also partially supported by the CUHK Teaching Development Grant.

7. References

[1] B. B. Kachru, *Asian Englishes: Beyond the Canon*. Hong Kong: Hong Kong University Press, 2005.

[2] K. Precoda, C. A. Halverson, and H. Franco, "Effects of speech recognition-based pronunciation feedback on second-language pronunciation ability," in *InSTILL*, 2000.

[3] A. Neri, C. Cucchiarini, and H. Strik, "ASR-based corrective feedback on pronunciation: does it really work?" in *INTERSPEECH*, Pittsburg, USA, 2006, pp. 1982–1985.

[4] J.-M. Kim, C. Wang, M. Peabody, and S. Seneff, "An interactive English pronunciation dictionary for Korean learners," in *INTERSPEECH*, 2004, pp. 1145–1148.

[5] F. Ehsani and E. Knodt, "Speech technology in computer-aided language learning," *Language Learning and Technology*, vol. 2, no. 1, pp. 45–60, July 1998.

[6] R. M. Kaplan and M. Kay, "Regular models of phonological rule systems," *Computational Linguistics*, vol. 20, no. 3, pp. 331–378, 1994.

[7] T. T. N. Hung, "Towards a phonology of Hong Kong English," *World Englishes*, vol. 19, no. 3, pp. 337–356, 2000.

[8] H. Meng, Y. Y. Lo, L. Wang, and W. Y. Lau, "Deriving salient learners' mispronunciations from cross-language phonological comparisons," in *ASRU*, 2007.

[9] 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 *INTERSPEECH*, 2008.

[10] C. Allauzen, M. Riley, B. Harb, J. Schalkwyk, M. Mohri, R. Sproat, and W. Skut, "OpenFST v1.1," http://www.openfst.org, 2009.

[11] D. Gildea and D. Jurafsky, "Automatic induction of finite state transducers for simple phonological rules," in *ACL*, 1995, pp. 9–15.

[12] D. Odden, *Introducing Phonology*. Cambridge, UK: Cambridge University Press, 2005.

[13] W. K. Lo, A. M. Harrison, H. Meng, and L. Wang, "Decision fusion for improving mispronunciation detection using language transfer knowledge and phoneme-dependent pronunciation scoring," in *ISCSLP*, 2008.