Improving mispronunciation detection and diagnosis of learners' speech with context-sensitive phonological rules based on language transfer

Alissa M. Harrison[1], *Wing Yiu Lau*[1], *Helen Meng*[1,2], *Lan Wang*[2]

[1]The Chinese University of Hong Kong, Shatin, N.T., Hong Kong
[2]CAS-CUHK Shenzhen Institute of Advanced Integration Technologies, Shenzhen, China
{alissa, wylau, hmmeng}@se.cuhk.edu.hk, lan.wang@siat.ac.cn

Abstract

This study demonstrates how knowledge of language transfer can enable a computer-assisted pronunciation teaching (CAPT) system to effectively detect and diagnose salient mispronunciations in second language learners' speech. Our approach uses a HMM-based speech recognizer with an *extended pronunciation lexicon* that includes both a model pronunciation for each word and common pronunciation variants of our target learners. The pronunciation variants in the extended pronunciation lexicon are generated based on language transfer theory (i.e knowledge from the first language is transferred to the second language). We find that a lexicon that characterizes language transfer using context-sensitive phonological rules can detect and diagnose errors better than a lexicon generated from context-insensitive rules. Furthermore, predicting errors from language transfer alone can approach the performance of a system where the lexicon is fully-informed of all possible pronunciation errors.

Index Terms: pronunciation training, mispronunciation detection, second language learning

1. Introduction

There has been considerable research on the requirements of effective computer-assisted pronunciation teaching (CAPT) software [1, 2]. An effective system should not only be able to detect mispronunciations but also provide corrective feedback which can help the learner rectify the error. The extensive review of CAPT by [3] finds that corrective feedback is crucial to CAPT and it "cannot rely on the student's own perception." The importance of corrective feedback has also been empirically demonstrated in the study of immigrant learners of Dutch [4].

Many of the existing approaches to developing CAPT software have focused on developing a numerical measure from the likelihood scores of an HMM-based speech recognizer to detect errors in the learner's speech [5, 6, 7]. While these systems have been able to develop pronunciation scores highly consistent with human ratings, they are inherently limited in the feedback that they can provide to the learner. These "goodness-of-pronunciation" scores can detect errors, but cannot diagnose the type of mispronunciation the learner has made. Without error diagnosis, learners may resort to trial-and-error to artificially improve their scores [8] .

An alternative approach to developing CAPT software is to incorporate linguistic knowledge of typical errors of the learners into the CAPT system. The feasibility of this approach has been demonstrated for Italian- and German-speaking learners of British English [9]. But there are still questions of how much linguistic knowledge of learners' errors is required to develop an effective system. This study attempts to address this issue and finds that predicting errors based on language transfer can enable a system to diagnose errors with a relatively high degree of success.

2. System Design

Figure 1 gives an overview of our general system design. The ASR utilizes an extended pronunciation lexicon, a grammar of fixed word order, and acoustic models trained on native speakers' speech data. The procedures for developing the extended pronunciation lexicon are discussed in the following section. The fixed word order grammar used by the recognizer mitigates the task of word recognition, and effectively reduces the problem to that of recognizing the pronunciation of a given word. The detection and diagnosis of the learners' mispronunciations is made possible by aligning the phone-level transcription of the recognizer with a model transcription based on native speaker pronunciations.

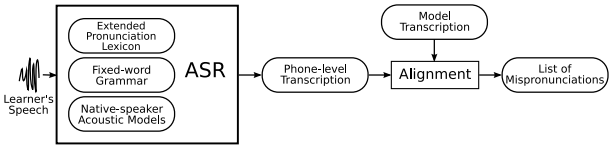

Figure 1: *Overview of ASR-based system to detect and diagnose second language learners' mispronunciations*

2.1. Speech Recognizer

The speech recognizer in our system uses cross-word triphone HMMs that contain 2000 states with 12 Gaussian mixtures per state. The implementation is based on the HTK Toolkit [11]. The acoustic models are trained on the TIMIT training set [12] which contains a total of 4620 sentences recorded by 462 speakers from eight dialect regions of the US.

2.2. Corpus design and annotation

The testing data of this paper comes from the CU Chinese Learners of English (CU-CHLOE) corpus used in [13]. We use 21 recordings of "The North Wind and the Sun" (9 male speakers, 12 female speakers). This piece is chosen because it exemplifies nearly all of the phonemes of English with the exception of the relatively rare /zh/ phoneme. Altogether, the passage is comprised of 113 words with a lexicon size of 64 words. The recordings were also annotated by a linguist using the Praat

annotation tool [14]. This human annotation is the "gold standard" (i.e. a transcription of the learners' actual speech) for our subsequent evaluation of the system.

3. Development of the Extended Pronunciation Lexicon

The extended pronunciation lexicon is key to our system's ability to detect and diagnose learners' mispronunciations. For each word in the extended pronunciation lexicon there is a model pronunciation (as determined by the TIMIT pronunciation dictionary) and additional *pronunciation variants* common to Cantonese learners of English. As outlined in Figure 2, we propose two methods for automatically generating these additional pronunciation variants based on a contrastive analysis of Cantonese and English: (1) context-insensitive and (2) context-sensitive phonological rules. These phonological rules are stated in the form of rewrite rules that can be applied to model pronunciations to generate the additional pronunciation variants.

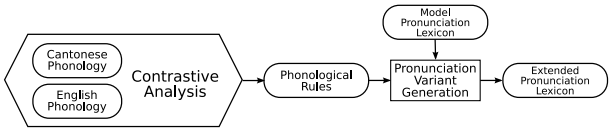

Figure 2: *Developing the extended pronunciation lexicon*

3.1. Contrastive analysis

Contrastive analysis is grounded in the theory of language transfer. The Contrastive Analysis Hypothesis states that sounds similar to the learner's first language will be easy for the learner to acquire while different sounds will present difficulty [10]. We conduct a contrastive analysis of Cantonese and English by examining the phonetic inventory and phonotactic constraints of the languages to determine phones and phone sequences present in English but lacking in Cantonese. Those phones which are not present in Cantonese are hypothesized to be substituted by Cantonese learners with phonetically-similar phones that do exist in Cantonese.

3.2. Context-insensitive phonological rules

From the above contrastive analysis we first developed a list of 43 context-insensitive rules [13] in the form of /α/ $\rightarrow$ /β/ (i.e. phone /α/ in the model pronunciation may be pronounced as /β/ by the learner). These 43 context-insensitive rules generated 2788 pronunciations variants in addition to the original 64 model pronunciations for the lexicon of "The North Wind and The Sun". These 2852 pronunciations make up the extended pronunciation lexicon which we will call Lexicon A.

We find two significant problems with using context-insensitive rules to generate an extended pronunciation lexicon: the lexicon grows exponentially and many pronunciations generated are rare or implausible in the learner's speech. For example, Cantonese does not have voiced stops (e.g. /b/, /d/, /g/) or consonant clusters (e.g. /s t r/) while English does. Cantonese learners may substitute voiceless counterparts (e.g. /p/, /t/, /k/) or delete consonants to cope with these difficult sounds. So our list must include rules like (1) /d/ $\rightarrow$ /t/, (2) /d/ $\rightarrow \emptyset$, and (3) /k/ $\rightarrow \emptyset$. Admittedly, we can see that rules (2) and (3) do not fully represent the knowledge gained from our contrastive analysis (i.e. deletion only occurs in consonant clusters). When these rules are applied to a word like 'could' /k uh d/, we generate pronunciation variants such as: /k uh t/, /uh d/, /uh/, etc. Note that while /k uh t/ is a plausible mispronunciation of 'could', the variants /uh d/ and /uh/ generated from (2) and (3) are so phonetically-distant from the model pronunciation of 'could' that they are considered implausible mispronunciations.

3.3. Context-sensitive phonological rules

To reduce the number of implausible pronunciations in the extended lexicon, context-sensitive rules were developed from the contrastive analysis of Section 3.1. The list of context-sensitive rules was compiled using the same list of context-insensitive rules but additionally specifying the phonetic environments that constrain its application. A total of 51 context-sensitive rules[1] were developed using the immediate neighboring segments and symbols for various linguistic classes: C for consonants, V for vowels, F for fricatives, and # for word-boundaries. These 51 context-sensitive rules generated 394 variants for this Lexicon B, significantly less than those generated by context-insensitive rules.

We can understand how context-sensitive rules solve the problem of over-generating implausible variants by reconsidering the variants /uh d/ and /uh/ generated by context-insensitive rules. These variants were generated because context-insensitive rules had no representational means to specify that deletion of consonants should only occur in consonant clusters. Context-sensitive rules solve this problem by allowing us to specify a phonetic environment that must be satisfied for the rule to apply. Thus, the consonant deletion rule from the previous section can be rewritten as /d/ $\rightarrow \emptyset$ / C _ where the left hand side specifies that /d/ must be preceded by a consonant in order for the rule to apply. When these context-sensitive versions of the previous rules are used to generate variants for a word like 'could', we see that the conditions of the deletion rules are not satisfied and thus implausible variants like /uh d/ and /uh/ are not generated.

3.4. Benchmark lexicons

To understand the effectiveness of the language transfer approach to automatic pronunciation generation, we also utilize two methods for manually generating pronunciations variants: (1) using an independent expert and (2) using the gold standard (see Section 2.2) to create a fully-informed list of the pronunciation variants.

A lexicon, Lexicon C, was manually-generated by an expert familiar with the common errors of Cantonese learners of English (different from the annotator of the test set). The pronunciation variants of Lexicon A were examined by the expert individually and those which seemed rare or implausible were removed. Pronunciations which the expert deemed likely but not in the lexicon were also added. This lexicon has a total of 361 pronunciations (64 model pronunciations, 297 variants).

Another benchmark lexicon, Lexicon D, was manually-generated by compiling all the unique pronunciations transcribed by the annotator in the test set. Lexicon D is said to be "fully-informed" because it includes all pronunciations attested in the test set, including those that may not be predicted by language transfer. It contains a total of 419 pronunciations (64 model pronunciations, 355 variants).

[1]The greater number of context-sensitive rules arose because some context-insensitive rules had more than one constraining phonetic environment and thus were written as multiple context-sensitive rules.

3.5. Evaluation procedures

The system was run on the 21-speaker test set using four different lexicons as described in the previous sections. The phone-level transcription output by the recognizer was aligned with the (1) model transcription derived from the TIMIT pronunciation lexicon and (2) gold standard as determined by the human annotator. The alignment was carried out through a bottom-up dynamic programming algorithm which returns the alignment that has the minimal sum-of-pairs score between the three phonetic transcriptions [15]. Substitutions were weighted with a cost of 10 and insertion / deletions had a cost of 7. Table 1 gives an example of a three-string alignment for the word 'north' where the minimal sum-of-pairs cost is 44.

Table 1: *A three-string phonetic alignment for the word 'north'*

MODEL:	n	ao	r	th
GOLD STANDARD:	l	ao		th
SYSTEM:	n	aa		th

This three-string alignment enables us to not only detect and diagnose mispronunciations in the learner's speech but also evaluate their accuracy as compared to the gold standard (i.e. human annotation).

4. Performance of Error Detection and Diagnosis

The performance of the system is measured in terms of its ability to (1) accurately detect correct and incorrect pronunciations of words, (2) accurately detect the correct and incorrect phones within a mispronounced word, and (3) accurately diagnose the errors in these phones. The first and second measures can be illustrated as a 2 x 2 classification matrix as shown in Figure 3.

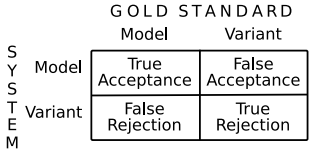

Figure 3: *Classification matrix for measuring accuracy of detecting correct and incorrect pronunciations*

At the word level, a true acceptance occurs when the learner's pronunciation is identical to the model pronunciation according to the gold standard and the system also recognizes the learner's pronunciation as equivalent to the model pronunciation. True rejection occurs when the pronunciation in the gold standard transcription differs from the model pronunciation and the recognizer also recognizes the pronunciation as one of the variants in the extended pronunciation lexicon (e.g. case of 'north' in Table 1). False rejection is where the system recognizes a pronunciation variant when the gold standard transcription is consistent with the model pronunciation, and vice versa with false acceptance. Higher rates of true acceptance and true rejection indicates better performance of the system. The Kappa coefficient is also given in the following tables to indicate the chance-corrected strength of agreement between the system and gold standard.

4.1. Detecting mispronounced words

We first evaluate the system's ability to detect which words are mispronounced by the learner. Table 2 shows the classification of words by the system as correct or incorrect compared to the gold standard. The columns titles TA, TR, FA, and FR are abbreviations for true acceptance, true rejection, false acceptance, and false rejection, respectively. The last column also gives the Kappa coefficient to indicate the strength of agreement between the gold standard and recognition transcription. The percentages are calculated according to the 2366 word tokens of the corpus.

We observe in Table 2 that a lexicon generated with context-sensitive rules (Lexicon B) leads to better detection of correct pronunciations (TA) but not necessarily better detection of mispronunciations (TR), as compared to context-insensitive rules (Lexicon A). Still, Lexicon B has better overall agreement with the gold standard as compared to Lexicon A due to the lower false rejection rate (FR). In the setting of language learning, we believe these attributes of Lexicon B are preferable (i.e. low false rejection rate and high true acceptance) as learners are apt to become frustrated with a system that falsely identifies correct pronunciations as incorrect [9]. Additionally, the accurate classification rate of Lexicon B (i.e. sum of TA and TR) is similar to the fully-informed benchmark Lexicon D (70.24% vs. 71.73%). This finding demonstrates that generating pronunciations variants based on language transfer alone (i.e. Lexicons A and B) is sufficient to obtain similar performance as lexicons which consider all possible causes of learner errors.

Table 2: *Classification of word pronunciations by the system*

Lexicon	TA	TR	FA	FR	Kappa
A (2366)	28.91%	39.73%	7.23%	24.13%	0.383
B (2366)	34.15%	36.09%	10.86%	18.89%	0.408
C (2366)	32.97%	36.31%	10.65%	20.08%	0.390
D (2366)	31.07%	40.66%	6.30%	21.98%	0.443

4.2. Detecting incorrect phones within mispronounced words

After detecting the mispronounced words, we evaluate how well the system can classify the correctness of the phones within these mispronounced words (i.e. subset of word tokens in TR and FR of Table 2). This measure is analogous to the previous except that we now consider phones instead of words. For example, consider the alignment given in Table 1: (1) the last phone of 'north' /th/ is identified by both the system and gold standard as /th/ so it is a case of true acceptance, (2) the third phone /r/ is a case of true rejection as both the system and gold standard agree that the learner deleted the /r/, (3) the first phone /n/ is a false acceptance by the system, (4) and the second phone /ao/ is a case of false rejection.

The results of Table 3 show that better classification performance at the phone-level is obtained with the lexicon generated with context-sensitive rules (Lexicon B) as compared to the one with context-insensitive rules (Lexicon A). Additionally, Lexicon B can accurately detect 72.68% (TA + TR) of the phones in mispronounced words. This number approaches the accurate classification rate of 74.92% in the expert-generated lexicon (Lexicon C) and 76.33% in the fully-informed benchmark lexicon (Lexicon D). Again, this demonstrates the power of predicting errors by language transfer alone.

Table 3: *Classification of phones by the system in words detected as mispronounced*

Lexicon	TA	TR	FA	FR	Kappa
A (5759)	50.69%	17.07%	6.58%	25.66%	0.302
B (5071)	55.31%	17.37%	7.38%	19.94%	0.373
C (5272)	57.15%	17.77%	6.32%	18.76%	0.417
D (5848)	56.58%	19.75%	4.14%	19.53%	0.467

4.3. Diagnosing mispronunciations

Based on the detection of incorrect phones within a mispronounced word, we evaluate the accuracy of the system in diagnosing the learner's mispronunciation. Table 4 shows the percentage of phones detected as mispronounced (i.e. phones in true rejection and false rejection columns of Table 3) that were transcribed identically between the system and human. The number in parentheses represents the total number of mispronounces phones for each lexicon.

Table 4: *Accuracy in diagnosis of phonetic mispronunciations*

Lexicon	Agreement
A (2461)	31.17%
B (1892)	42.71%
C (1926)	37.69%
D (2297)	46.15%

We find Lexicon B has significantly better diagnostic performance than Lexicon A (42.71% vs. 31.17%) due in part to its lower false rejection rate. It also performs better than Lexicon C despite the latter having better rates of correct and incorrect phone detection. The error agreement rate of Lexicon B approaches the upper-bound performance as determined by Lexicon D. Although context-sensitive rules have been shown to lead to much better performance in error diagnosis, we acknowledge the upper-bound of the system (46.15%) is lower than desired. Since Lexicon D is fully-informed with all variants from the gold-standard, we believe the relatively low agreement is due to poor discrimination of similar phones by the acoustic models in the speech recognizer.

5. Conclusions and Future Work

ASR technology in CAPT systems have a lot to offer the language learning community, but ASR-based tools must be designed such that they can provide corrective feedback to the learner. In this paper, we have proposed a system design that can not only detect mispronounced words but also correctly diagnose the type of errors made by the learner. The error diagnosis capability of our system is especially crucial as it can be used to develop detailed corrective feedback for the learner.

Our system evaluation has shown that generating a pronunciation lexicon with context-sensitive rules has better performance than a lexicon generated with context-insensitive rules. We have also demonstrated that a lexicon generated from context-sensitive rules can detect and diagnose mispronunciations at a rates comparable to manually-generated lexicons. These results show that predicting errors through contrastive analysis alone is sufficient to enable detection of the majority of learners' errors. This is significant as the contrastive analysis procedure can be carried out on any pair of languages and, for many well-known languages, using existing linguistic studies. Secondly, this method of generating pronunciation variants with context-sensitive rules does not necessarily depend on corpus data. Thus, our system design can readily be utilized for learners from first and second language backgrounds different than those of this study.

We believe our approach is a promising direction for developing CAPT tools and see the potential for further performance improvements via discriminative training and pronunciation scoring. Discriminative training techniques may be applied to improve the ability of the system to distinguish phonetically-similar phones. Additionally, while context-sensitive rules have high performance in diagnosing errors, they may benefit from better detection of mispronunciations. Previous studies, as mentioned in Section 1, have shown the ability of "goodness-of-pronunciation score" to detect errors. Thus, we believe this metric can combine with our context-sensitive phonological rules approach to further improve the performance of mispronunciation detection.

6. Acknowledgements

This work is partially supported by the CUHK Teaching Development Grant. The authors would like to thank Ms. Pauline Lee of the Independent Learning Center at CUHK for her helpful feedback.

7. References

[1] Pennington, M. C., "Computer-aided pronunciation pedagogy: Promise, limitations, directions," CALICO, 12(5), 427-440, 1999.

[2] Wachowicz, K. A. and Scott, B., "Software that listens: Its not a question of whether, its a question of how," CALICO, 16(3), 253-276, 1999.

[3] Ehsani, F. and Knodt, E., "Speech technology in computer-aided language learning", Lg Learning and Tech, 2(1), 45-60, 1998.

[4] Neri, A., Cucchiarini, C., and Strik, H., "ASR-based corrective feedback on pronunciation: does it really work?" Proc. Interspeech-2006, 1982-1985.

[5] Witt, S. M. "Use of speech recognition in computer-assisted language learning," Ph.D. diss., Cambridge Unviersity, 1999.

[6] Kim, I.-S. "Automatic speech recognition: Reliability and pedagogical implications for teaching pronunciation," Educational Technology and Society, 9(1), 322-334, 2006.

[7] Franco, H., Abrash, V., Precoda, K., Bratt, H., Rao, R., Butzberger, J., Rossier, R. and Cesari, F., "The SRI Eduspeak system: Recognition and pronunciation scoring for language learning," Proc. InSTIL-2000, 123-128.

[8] Eskenazi, M. and Hansma, S., "The fluency pronunciation trainer," Proc. STiLL-1998.

[9] Howarth, P. Pezzota, D., Galbiati, R., and Bisiani, R., "Validation report," ISLE, Tech. Rep., 2000.

[10] Lado, R., Linguistics Across Cultures: Applied Linguistics for Language Teachers, University of Michigan, Ann Arbor, 1997.

[11] Young, S. J. Odell, J. Ollason, D., and Woodland, P.C., "The HTK book Entropic", Cambridge Research Laboratory, 1996.

[12] Fisher, W., Zue, V., Bernstein, J., and Pallet, D., "An Acoustic-Phonetic Data Base", J. Acoust. Soc. Am. 81, Suppl 1, 1987.

[13] Meng, H. M. Lo, Y. Y., Wang, L., and Lau, W.Y., "Deriving Salient Learners' Mispronunciations from Cross-Language Phonological Comparisons" Proc. ASRU-2007.

[14] Praat: doing phonetics by computer, Paul Boersma and David Weenink, Institute of Phonetic Sciences, University of Amsterdam. http://www.fon.hum.uva.nl/praat/

[15] Gusfield, D., Algorithms on strings, trees, and sequences: computer science and computational biology, CUP, NY, 1997.