Automatic Derivation of Phonological Rules for Mispronunciation Detection in a Computer-Assisted Pronunciation Training System

Wai-Kit Lo, Shuang Zhang and Helen Meng[1]

Department of Systems Engineering and Engineering Management,
Shun Hing Institute of Advanced Engineering,
The Chinese University of Hong Kong

{wklo, zhangs, hmmeng}@se.cuhk.edu.hk

Abstract

Computer-Assisted Pronunciation Training System (CAPT) has become an important learning aid in second language (L2) learning. Our approach to CAPT is based on the use of phonological rules to capture language transfer effects that may cause mispronunciations. This paper presents an approach for automatic derivation of phonological rules from L2 speech. The rules are used to generate an extended recognition network (ERN) that captures the canonical pronunciations of words, as well as the possible mispronunciations. The ERN is used with automatic speech recognition for mispronunciation detection. Experimentation with an L2 speech corpus that contains recordings from 100 speakers aims to compare the automatically derived rules with manually authored rules. Comparable performance is achieved in mispronunciation detection (i.e. telling which phone is wrong). The automatically derived rules also offer improved performance in diagnostic accuracy (i.e. identify how the phone is wrong).

Index Terms: pronunciation training, phonological rules, mispronunciation detection, language learning

1. Introduction

The growing number of second language (L2) learners worldwide creates an increasing demand for language learning resources. It is estimated [1] that the number of English learners in India and China alone is 533 million, which is greater than the combined population of the USA, UK and Canada. It is expected that there will be a shortage of qualified teachers to satisfy such language learning needs. Computer-Assisted Pronunciation Training (CAPT) is a viable supplementary solution for the limited resources.

CAPT systems can use automatic speech recognition (ASR) technology to offer productive training for language learners [2][3]. A major advantage of using CAPT is that the learners can practice at anytime at their convenience before getting help from a human teacher. It can also provide a personalized learning environment to reduce the learners' anxiety, as well as provide consistent, objective and individualized feedback.

Our approach to CAPT is to prompt the learner to read pre-designed materials. The system then performs mispronunciation detection and diagnosis on the recorded utterances. Diagnostic feedback is returned to the learners pinpointing any problems in their pronunciations. We focus on mispronunciations due to language transfer effects [6][7][8]. Language learners have a tendency to substitute phones in their L2 speech with phones in their primary languages (L1). Such language transfer effects may sometimes be perceived as accents in L2 and add "color" and "texture" to the non-native language. However, there are also cases of negative language transfer, where the phonetic substitutions lead to mispronunciations. To identify these pronunciation problems, we have designed phonological rules manually to capture negative language transfer effects [6][7]. However, preparation of such rules requires linguistic expertise in both L1 and L2. The rule set created depends on the author's (or linguist's) experience and hence may not guarantee good coverage of possible observations. Also there is no guarantee for consistency among different authors. Hence we propose an automatic method to derive phonological rules directly from L2 data. This can save much human effort and can be easily portable to any pair of L1 and L2.

2. Capturing Negative Language Transfer Effects with Phonological Rules

Our work focuses on pronunciation training for native Cantonese speakers learning English. There are significant phonological differences between Cantonese and English. It is believed that L2 learners will apply the phonological characteristics of their L1 for the L2 [4]. For example, all plosives and fricatives in Cantonese are unvoiced, plosives in coda positions are always unreleased, etc. The L2 speech of a native Cantonese speaker often substitutes the voiced fricative /v/ with an unvoiced fricative /f/. Hence, one may design the phonological rule (/v/ → /f/) to capture this particular phenomenon. There are also other phenomena specific to the Cantonese-English language pair. Application of a set of carefully designed phonological rules in CAPT [6][7][8] enables our system to detect mispronunciations as well as generate corrective feedback for the learners. (e.g., indicating that the phone /v/ is mispronounced as /f/, which is a case of devoicing).

2.1. Context-sensitive phonological rules

Different phonetic confusions may be realized in different phonetic contexts. For example, while the confusion (/v/ → /f/) occurs in both word-initial and word-final positions, there is also the confusion (of /v/ → /w/) which occurs mainly in word-initial positions. This motivates the design of context-sensitive phonological rules, as illustrated in Figure 1.

/v/→ /f/

(a)

/v/→ /f/ | _ #
/v/→ /w/ | # _
/v/→ /f/ | # _

(b)

Figure 1. Example (a) illustrates a context-free phonological rule and (b) shows context-sensitive rules. The symbols on left-hand-side, right-hand-side and context are in DARPABET. '#' indicates a word boundary.

We adopted the rule format

$$\varphi \rightarrow \psi / \lambda \_ \rho,$$

[1] Corresponding author

which denotes that phone φ may be substituted by the phone ψ when it is preceded by the phone λ and followed by the phone ρ. Similarly, by including the null symbol ε, we can encode a phone insertion by $\varepsilon \rightarrow \psi$ and phone deletion rule by $\varphi \rightarrow \varepsilon$.

2.2. Extended Recognition Network (ERN)

Given the canonical pronunciation of a word (e.g., /n ao r th/ for "north"), we can apply the phonological rules to obtain a list of possible mispronunciations. These mispronunciations can be represented in an ERN, which is a compact representation of the canonical pronunciations, as well as its expansion to possible mispronunciations [8]. This compact representation saves storage and reduces computation in an ASR by avoiding searches in redundant phone paths. By representing the phonological rules as a collection of finite state transducers, we can easily generate the compact ERN from any canonical pronunciations. Figure 2 shows an example for the ERN of the word "north", containing both the canonical pronunciation (/n ao r th/) and the pronunciation variants predicted by the phonological rules, e.g., /n ao r f/, /n ow th/, etc. Hence the ASR can select the acoustically best matching phone sequence as the detection output.

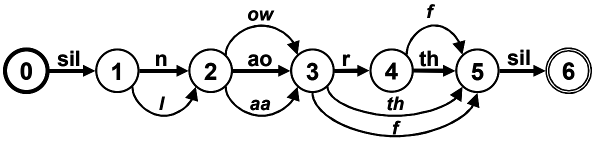

Figure 2. ERN is a compact representation of pronunciation variants in the form of a network. It includes the canonical pronunciations (the bolded path in the middle) and all variants covered by the phonological rules.

3. Automatic, data-driven derivation of phonological rules

We devised an automatic, data-driven method to derive phonological rules from L2 speech data. This method makes use of phonetically transcribed L2 speech data, together with canonical pronunciations. Our approach is based on a few assumptions: (i) differences in the phonetic transcriptions and the canonical pronunciations are due to negative language transfer effects; (ii) other interferences such as misread prompts, unknown words, transcription errors, ambiguity due to multiple accented pronunciations, disfluencies in spontaneous speech etc., do not dominate; (iii) the pronunciation dictionary provide good coverage of canonical pronunciations of all words. The proposed approach is summarized as shown in Figure 3. Detailed descriptions and results of empirical evaluation will be given in the following sub-sections.

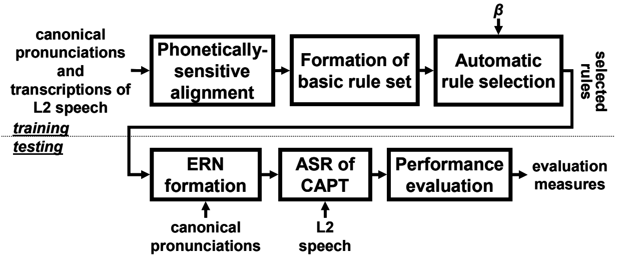

Figure 3. Schematic diagram of an automatic, data-driven derivation method for phonological rules in a CAPT system.

3.1. Corpus and rule selection criterion

The experiments are carried out using the Cantonese part of the CU-CHLOE corpus. This part of the corpus contains prompted speech data collected from 100 native Cantonese speakers (50 male and 50 female) reading several kinds of carefully designed materials: (i) the Aesop's Fable "The North Wind and the Sun" (6 utterances), (ii) phonemic sentences (20 utterances), confusable words (10 utterances), and minimal pairs (50 utterances). All speech data are manually transcribed and the canonical pronunciations of all words can be readily obtained from electronic dictionaries (e.g., TIMIT, CMUDict, etc.). In this work, the 100 speakers are divided into disjoint training (25 male and 25 female) and test (25 male and 25 female) sets. Evaluation is based on the F1-score (see Equation 1) [1]. This is computed from the precision and recall of mispronunciations captured by the phonological rules, with reference to the actual ones observed in from the data.

$$F1 = \frac{2\,P\,R}{P\;+\;R} \tag{1}$$

where P is the precision and R is the recall.

3.2. Formation of the basic rule set

The automatic, data-driven derivation of phonological rules is based on a process of generation and selection. First, the canonical pronunciations are aligned with the manual transcriptions of L2 speech using phonetically-sensitive alignment [8]. An example is shown in Figure 4.

	left context		right context	occurrence count
canonical	#	dh	ax	471
manual	#	d	ax	

Figure 4. An aligned pair of canonical and manual transcriptions of L2 speech, for the word "the".

From the aligned phones, we extract all mismatched phone pairs together with their left and right contextual phones. This generates a set of context-sensitive phonological rules, called the *basic rule set*. For instance, the mismatched phone pair in Figure 4 will give us the context-sensitive rule "dh → d / # _ ax" (voiced inter-dental fricative "/dh/ is substituted by a voiced alveolar plosive /d/ when it is in the word-initial position and followed by an /ax/"). In this way, we obtain 2,320 rules from the aligned training data as the *basic rule set*. This set is guaranteed to provide 100% coverage of the mispronunciations in the training data.

3.3. Top-down rule selection approach

The 2,320 context-sensitive phonological rules in the basic rule set provide full coverage of all mispronunciations in the training set, including those with rare occurrences. These tend to lead to false alarms in mispronunciation detection. Hence, we devise a rule selection procedure to obtain a reduced set of rules and maintain good coverage of the observed mispronunciations. We first rank the 2,320 rules in descending order of occurrence frequency, as shown in Figure 5a. It can be seen that many rules have low occurrences. A closer look into the data reveals that some of them are due to misreading by the subjects while others are due to guessed pronunciations for words unfamiliar to the speakers.

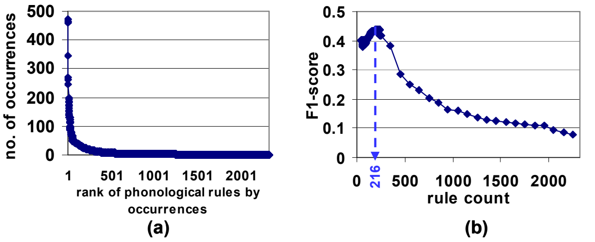

Figure 5. (a) Ranking of context-sensitive phonological rules (in the basic rule set) by occurrence counts. (b) F1-score based on the selected top-N rules (optimal N = 216).

The automatic phonological rule derivation method incrementally builds the selected rule set. Rule consideration follows the descending order of occurrence frequency observed in the training data. We add rules one-by-one to the selected rule set. With each addition, we evaluate the mispronunciation retrieval performance of the selected rule set by means of the F1-score (see Figure 5b). It can be seen that the F1-score peaks at 216 rules.

3.4. Evaluation of the selected rule set on mispronunciation coverage

We have applied the selected 216 rules to the disjoint test set. The mispronunciation detection performance is shown in Table 1, and is compared with the results obtained using the manually authored rule set reported in our previous work [8]. Both sets of rules have the same rule format. Analysis shows that while the automatically selected rule set (216 rules) contains more rules than the manually authored rules set (with 51 rules), the former gives improved precision and recall and a 93.2% relative improvement in F1-score.

Rule set	Manually authored (51 rules)	Automatically derived (216 rules optimized from the training data)	
Evaluation data	Test set	Training set	Test set
Hits	9,131	12,990	12,544
False alarms	54,773	26,704	27,095
Precision	0.1429	0.3273	0.3165
Recall	0.4561	0.6612	0.6265
F1-score	0.2176	0.4378	0.4205

Table 1. Comparison between manually authored (51 rules) and automatically selected phonological (216 rules) rule sets, in capturing mispronunciations observed in the test set.

v → f / # _ ay	d → 0 / axr _ #
v → f / ey _ #	d → 0 / ay _ #
v → f / ax _ #	d → 0 / ih _ #
v → f / ih _ #	d → 0 / ix _ #
v → f / ae _ el	d → 0 / iy _ #
v → f / iy _ #	d → 0 / n _ #
⋮	d → 0 / n _ z
z → s / ow _ #	⋮
z → s / uw _ #	axr → ax / aw _ d
z → s / ey _ #	axr → ax / d _ d
z → s / ae _ #	axr → ax / v _ m
z → s / ax _ #	axr → ax / v _ #
⋮	axr → ax / f _ #
d → t / uw _ #	axr → ax / iy _ #
d → t / ae _ #	axr → ax / dh _ z
d → t / jh _ #	⋮
d → t / aa _ #	r → ax / ae _ #
d → t / eh _ #	r → ax / ao _ #
d → t / ix _ #	⋮
d → t / n _ #	th → f / # _ ih
d → t / r _ #	th → f / r _ #
⋮	⋮

Figure 6. Excerpt of automatically derived phonological rules by the proposed approach.

Figure 6 shows examples of the automatically selected rules. It can be seen that the automatic method has successfully extracted many of the common mispronunciations made by the Cantonese speakers. Mispronunciations captured in these example rules can be roughly summarized as:

- devoicing the fricative /v/ as /f/,
- devoicing the fricative /z/ as /s/,
- devoicing and substituting the inter-dental fricative /th/ with /f/,
- devoicing and substituting the voiced plosive /d/ with /t/,
- deletion of retroflex in /axr/, and
- substituting the retroflex /r/ by the reduced vowel /ax/.

4. Mispronunciation detection with generalized rule derivation and ASR

The 216 rules selected in the previous section were optimized for the F1-score. This puts equal emphasis on precision and recall. It does not take into consideration of the effect of the ASR in the mispronunciation detection process. In a practical task, one may want greater control of the weighting between precision and recall. Considering F1-scores alone may not lead to the optimal operating point when ASR is involved. In the ideal case of a CAPT system with a "perfect" ASR engine, having a huge list of candidate mispronunciations in the ERN (e.g., including those that do not appear frequently) does not hurt. It is because a "perfect" ASR can always pick out the appropriate pronunciation in the input speech (either correct or mispronounced), provided that the pronunciation is present in the recognition network. This suggests that a higher recall in the F1-score is more desirable for a "perfect" ASR. However, the optimal operating point in a practical CAPT system involves a trade-off among the combined effects of precision, recall and recognition accuracy. In other words, for an imperfect ASR, the best weighting between precision and recall is unknown. We will present our investigation in this section.

4.1. The generalized F_β-measure

We performed the automatic rule derivation process with the generalized F_β-measure (Equation 2). Higher values of β emphasize recall. Hence, a larger β will lead to a larger rule set (see Figure 7).

$$F_\beta = \frac{(1+\beta^2)\,P\,R}{\beta^2\,P\,+\,R} \qquad (2)$$

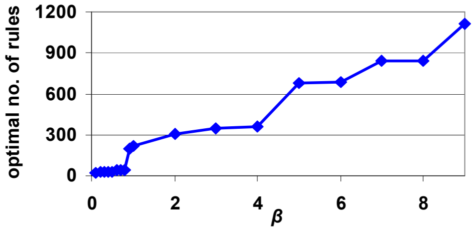

Figure 7. Optimized number of rules selected based on the F_β-measure at various values of β.

4.2. Mispronunciation detection with ASR at different values of β

With the optimized rule set generated for a specific value of β, we apply the rules to generate an ERN for every word in the prompting text. Mispronunciation detection is then carried out by performing a forced-alignment of the speech data with the ERNs.

We apply each rule in the selected set independently on to the canonical pronunciation(s) of the word in the prompt. An intermediate output is formed by taking the union of the outputs from all rules. Then all rules are re-applied to this intermediate output to obtain the final output. This two-pass rule application procedure covers the sequential effect in the phonological processes that lead to mispronunciations.

Figure 8 shows the ASR performance in terms of the percentage of matching phones between the manual transcription and recognition output for different values of β. The match is between the recognition output and the manual transcribed phones from the non-native speaker input speech, which includes mispronunciations. Percentage of match (indicated as the usual ASR phone correctness in Figure 8) obtained using our ASR and the basic rule set (2,320 rules) is 54.86%. Significant improvements over the basic rule set are obtained by using the optimized rule sets at different values of

β. Note that increasing β values lead to larger rule sets and bushier ERNs which include more possible mispronunciations. Without a "perfect" ASR, this can introduce a higher degree of phonetic confusion and hence the phone correctness declines.

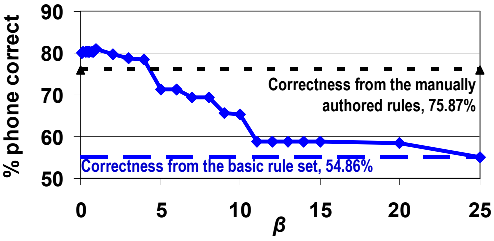

Figure 8. Phone correctness (indicating the percentage of phone match between the manual transcriptions and recognition results) of the CAPT in mispronunciation detection when using ERNs generated from the rule sets derived at various values of β. When the manually authored rule set is applied, the phone correctness is 75.87%. When the basic rule set is applied, the phone correctness is 54.86% (when β reaches 25).

Performance on mispronunciation detection (FAR and FRR) using ERNs for different values of β is shown in Figure 9 (remark: β=1 corresponds to F1 results in Section 3.3). Bushier ERNs have better coverage of possible mispronunciations, which lead to lower FAR values (more errors detected). On the other hand, there is also greater phonetic confusion, which leads to higher FRR values (more correct pronunciations rejected). Figure 10 shows the Diagnostic Accuracies (DA), which measures the correctness in identifying the type of mispronunciations (i.e., identifying $\varphi \rightarrow \psi$, and ψ equals ε for deletions). A higher DA implies that the system generates more accurate feedback for the users.

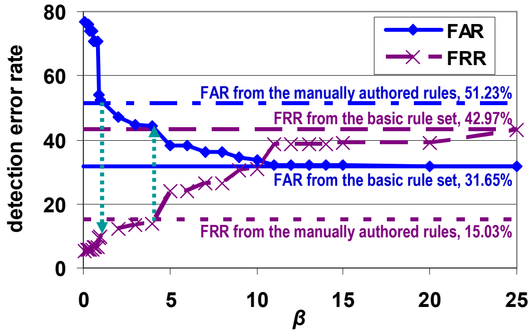

Figure 9. Mispronunciation detection performance based on ERNs obtained from rule sets derived at different β. Applying the manually authored rule set from [8] achieved an FAR of 51.23% and an FRR of 15.03%. Applying the basic rule set achieved an FAR of 31.65% and an FRR of 42.97%. The curves of FAR and FRR asymptotically approach these results from the basic rule set when all rules are included in the selected rule set. Equal Error Rate (EER) is about 33%.

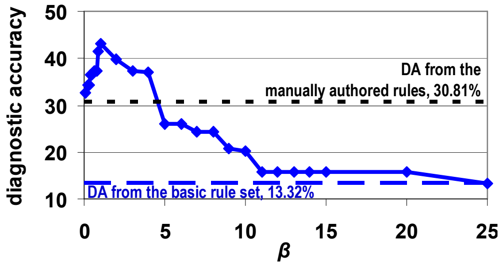

Figure 10. Diagnostic Accuracy (DA) of ERNs obtained at different β values. Applying the manually authored rule set from [8] achieved 30.81%. Applying the basic rule set achieved 13.32%.

We summarize our experimental findings as follows:

- The automatic, data-driven method of phonological rule derivation allows us to choose a desirable operating point (combination of FAR and FRR).
- In comparison with the manually authored rule set [8] (FAR=51.23% and FRR=15.03%), the automatically derived rules achieve comparable performance as indicated by the 2 dotted arrows in Figure 9.
- In terms of Diagnostic Accuracy, the automatically derived rules achieve better performance, when compared with the manually authored rule set (for $\beta < 5$).
- When $1 < \beta < 5$, the automatically derived rules achieve better performance than the manually authored rule set [8] in all three evaluation metrics (FAR, FRR and DA).
- Although the speakers in the training and test sets are disjoint, the text prompts for recording are the same. This means that the training set provides full lexical knowledge of the test set in terms of canonical pronunciations. However, the training set does not offer any knowledge about mispronunciations made by speakers in the test set.

5. Conclusions

We present an automatic, data-driven approach for phonological rule derivation to capture possible mispronunciations in L2 speech. Experimental results from an L2 corpus of 100 speakers (Cantonese learners of English) indicate that mispronunciation detection performance of the automatically derived rules is comparable to the set of manually authored rules. The automatic approach enhances the portability of our CAPT system to other language pairs, e.g., Mandarin and English. It also offers the choice of a desirable operating point (combination of FAR and FRR) for mispronunciation detection, based on the needs of the CAPT application. As the ASR performance improves, the approach may be enhanced by the using recognition networks (ERNs) with higher complexities, because these have greater coverage of possible mispronunciations.

6. Acknowledgements

This work is affiliated with the CUHK MoE-Microsoft Key Laboratory of Human-centric Computing and Interface Technologies. This project is partially supported by a grant from the HKSAR Government Innovation and Technology Fund (ITS/286/09). We wish to thank Alissa Harrison for her support, especially in authoring the phonological rules. We also wish to thank the labelers for their transcriptions.

7. References

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

[2] G. Kawai and K. Hirose, "A Call System Using Speech Recognition to Teach the Pronunciation of Japanese Tokushuhaku," in *Proc. of STiLL1998*, pp. 73-76, 1998.

[3] T. Kawahara *et al.*, "Practical Use of English Pronunciation System for Japanese Students in the CALL Classroom," in *Proc. of INTERSPEECH2004*, pp. 1689-1692, 2004.

[4] R. Lado, *Linguistics Across Cultures*. University of Michigan Press. 1957.

[5] C. J. van Rijsbergen, *Information Retrieval*, Butterworth, 1979.

[6] H. Meng, Y. Y. Lo, L. Wang, and W. Y. Lau, "Deriving Salient Learners' Mispronunciations from Cross-language Phonological Comparisons," in *Proc. of ASRU2007*.

[7] 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 *Proc. of INTERSPEECH2008*.

[8] A. M. Harrison, W. K. Lo, X. J. Qian, and H. Meng, "Implementation of an Extended Recognition Network for Mispronunciation Detection and Diagnosis in Computer-Assisted Pronunciation Training," in *Proc. of SLaTE2009*.