A Two-Stage Mispronunciation Detection Approach for Computer-Assisted Pronunciation Training

Hua Yuan[1], Junhong Zhao[2], and Jia Liu[1]
[1]Tsinghua National Laboratory for Information Science and Technology,
Department of Electronic Engineering, Tsinghua University, Beijing 100084
[2]State Key Laboratory on Transducing Technology, Institute of Electronics,
Chinese Academy of Sciences, Beijing 10008
yuanh08@mails.tsinghua.edu.cn, zhaojunhong09@mails.gucas.ac.cn, liuj@tsinghua.edu.cn

***Abstract*—In this paper, we propose a two-stage mispronunciation detection approach for computer-assisted pronunciation training. In the first stage, the selected phonological rules are used to cooperate with ASR to detect mispronunciations based on language transfer. Because the first stage detection can only deal with the pronunciation errors in the scope of the phonological rules, and detection performance is depressed with the imperfect phoneme acoustic model. The rescoring method based on duration normalized log posterior probability (NLPP) is employed in the second stage to identify the recognition speech unit again. Furthermore, a new F_α-score ranking criterion is proposed for the first stage to balance the mispronunciation coverage and recognition confusion, in the aim of minimizing the cost of total detection errors. The experiment shows that the method only with phonological rules gets a best performance of 19991 total detection errors, and the normalized log posterior probability method costs 22264 total errors. Finally, the two-stage detection approach can reduce the total errors to 19498.**

I. Introduction

The Computer-Assisted Pronunciation Training (CAPT) system is designed to help the native learners learn a foreign language on the pronunciation problem [1][2]. It's a common experience that learners often "transfer" elements of their native language onto the speech patterns of the target language. The previous work in [3][4][5][6] is dedicated to the mispronunciation detection of Cantonese speakers learning English. The automatic derivation of phonologic rules approach cooperating with the automatic speech recognition (ASR) technology is put forward in the CAPT to capture the mispronunciations induced by the negative language transfer. The promising result of these work indicates the significant value of the phonological rule method, and motivates us to investigate the application of this method in language study further.

This work attempts to design a two-stage mispronunciation detection approach. In the first stage, a candidate rule set is extracted from the training data. Phonological rules are selected according to the ranking result of this candidate rule set to model the mispronunciations. Apparently the ranking metric which is used to rank the candidate rules is crucial in this stage. In the method [6], the rules were ranked by the trigger count, so the selected rules heavily depend on exist errors in given data: some rules may bring much more false alarms than hits, which may bring large confusion to the extended recognition network. So here in order minimize this effect, we use the F_α-score criterion to rank the rules before selecting the top-N rules to achieve good balance of the mispronunciation coverage and recognition confusion. With the proposed ranking metric we can seek a best performance in mispronunciation detection with lowest cost in total errors, which means the system can maximize pronunciation error detection while minimizing the discouragement to learners.

The phonological rule method relies on the selected phonological rules. Therefore firstly it can't cover the errors which are beyond the scope of the rule set. Secondly, the phoneme acoustic model in ASR is imperfect enough to choose the closest pronunciation to the speech segment. On the other hand, the acoustic likelihood score is usually used to measure the similarity between speech segment and corresponding speech unit [7][8], and it performs well in judging whether the pronunciation is correct [9]. However, due to the constraints caused by the given speech unit, the acoustic likelihood score method is only useful for detecting replacement mispronunciation. So in the second stage, we introduce the acoustic likelihood score based on duration normalized log posterior probability (NLPP) as the discriminant feature to further distinguish correct pronunciation from the error ones, on the basis of the recognition results of first stage.

The following of this paper is organized as follows. The mispronunciation detection method with phonological rules is described in Section II. The two-stage detection method is showed in Section III. Section IV gives the experiment results. And conclusions are draw in Section V.

II. Mispronunciation Detection with Phonological Rules

This work is supported by the NSFC and RGC scheme (Project 60931160443 and N_CUHK414/09), National High Technology Research and Development Program of China (Project 2008AA040201), and National Natural Science Foundation of China (Project 90920302).

The phonological rules are used to model the mispronunciations from the data. Firstly the phonological rules are extracted from the alignment between manual transcription and pre-provided canonical pronunciation transcription. These rules form the basic rule set. Then the rules are ranked. After that, the N-best rules are selected and applied to generate the extended recognition network for mispronunciation detection. The process of mispronunciation detection with the selected rules is summarized as follows: 1) use finite state transducers (FSTs) to represent the rules; 2) generate extended recognition network (ERN) for a sentence by composing the canonical pronunciation with the rule FSTs; 3) detect the pronunciations of the learner's speech by recognition using the corresponding ERN as recognition grammar; 4) align the detected pronunciation with the canonical pronunciation; 5) identify mispronunciation from the mismatch in the aligned pronunciations and provide diagnostic feedback.

Here we illustrate the three steps of the generation of top-N rules in detail.

A. *The extraction of rules*

The canonical transcription of the training set is aligned with the manual transcription using a phonologically-sensitive string alignment [5]. From the aligned phoneme pairs, the phonological rules are extracted as the basic rule set, in the following form:

$$\phi \to \varphi / \lambda \_ \rho , \qquad (1)$$

which means the target phoneme ϕ is replaced by the phoneme φ, if its preceding phoneme is λ and following phoneme is ρ. The rule represents an insertion mispronunciation when ϕ is null, a deletion mispronunciation in the case of that φ is null, and a substitution one if ϕ is a phoneme different from φ.

B. *The ranking of rules*

In this work, the rules are ranked by the following criterion respectively:

Trigger

The trigger criterion is used in [6], which means the triggered mispronunciation count of each rule in the data. The triggered mispronunciation refers to the actual ones observed from the data.

F_α **-score**

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

where P denotes the precision of the current rule, R denotes the recall of the current rule, and α is a parameter to balance the precision and recall. Just like the trigger criterion, the computation of P and R refers to the actual observation ones.

Each applied rule will bring false alarms at the same time as hits. In order to detect mispronunciations as more as possible with false alarms as less as possible, we use the parameter α to balance the triggering mispronunciation and confusion to the recognition network of each rule.

New F_α **-score**

The computation of new F_α -score is the same as F_α -score, except precision and recall are obtained from the results of recognition, not directly from the data.

Because the phonological rules will be used cooperating with the ASR, the final recognized hits and false alarms may differ greatly from the statistic value directly from the data. So in the new F_α -score criterion, we first apply all the rules of basic rule set to detect the mispronunciation on the training data, and get the count of hits and false alarms from the detection results. Then the precision and recall can be computed according to the hits and false alarms.

C. *The selection of rules*

We select the top-N rules in the ranking rule set as the optimal rule set.

Then the generation flow of optimal rule set can be described as Fig. 1.

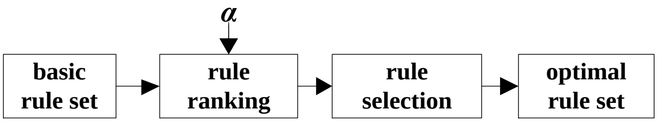

Fig. 1. The diagram of rule generation flow.

III. TWO-STAGE MISPRONUNCIATION DETECTION

A. *Normalized log posterior probability (NLPP)*

As the acoustic likelihood score, the log posterior probability represents the degree of similarity between speech segment and speech unit. Given the observation O of a speech segment and the corresponding speech unit q (the monophone model), LPP is computed as:

$$\begin{aligned} LPP(q \mid O) &= \log p(q \mid O) = \log\left(\frac{p(O \mid q) \cdot P(q)}{\sum_{q_i \in Q} p(O \mid q_i) \cdot P(q_i)}\right) \\ &\approx \log\left(\frac{p(O \mid q)}{\max_{q_i \in Q} p(O \mid q_i)}\right) \end{aligned}, \qquad (3)$$

where Q is the speech unit set.

Since the triphone model can describe the co-articulation better than the monophone model, we use the triphone model to compute the LPP, and then the calculation of LPP becomes complex. We modify (3) to the triphone computation form:

$$\begin{aligned} LPP(x-q+y \mid O) &= \log p(x-q+y \mid O) \\ &\approx \log\left(\frac{p(O \mid x-q+y)}{\max_{q_i \in Q} p(O \mid x-q_i+y)}\right) \end{aligned}, \qquad (4)$$

where x is the previous phoneme of q, y is the following phoneme of q, and $x-q+y$ represents a triphone model.

Furthermore, we use the duration d of speech segment to normalize the LPP, so the duration normalized log posterior probability (NLPP) is computed as:

$$NLPP(x-q+y|O)=\frac{1}{d}\log p(x-q+y|O) \approx \frac{1}{d}\log(\frac{p(O|x-q+y)}{\max_{q_i\in Q} p(O|x-q_i+y)}) \quad . \tag{5}$$

B. Two-stage mispronunciation detection

The mispronunciation prediction method can identify the errors, but performs invalid in these cases: 1) the mispronunciation doesn't appear in the training data, or the rule of the mispronunciation is pruned; 2) The recognizer with the imperfect acoustic model can't pick up the right pronunciation of speech utterance. In contrast, the NLPP depends on the pronunciation transcription, and is able to deal with any kinds of mispronunciation for the given speech segment and corresponding pronunciation transcription. So when there is an unknown insertion error (meaning unknown transcription), the NLPP is out of work.

The two-stage mispronunciation detection is designed to combine the strengths of mispronunciation and LPP. In the first stage, we take the mispronunciation prediction as the baseline, to get the fundamental detection result and speech segment for each pronunciation unit. Then we compute the NLPP for each speech segment and give the final diagnostic results. The final flow of the detection is described as Fig. 2.

IV. Experiment

A. Experiment setup

In order to verify the performance of the proposed method, we conduct experiments on the CU-CHLOE, which is designed to capture pronunciation variants in Chinese learner's English speech. It consists of speech data from 111 Mandarin speakers and 100 Cantonese speakers. The reading materials are designed the same as [6]. There are 86 utterances for each speaker. In this work, we make use of the Mandarin part of the corpus. 30 male and 25 female speakers are selected as the training set while the remaining 31 male and 25 female are used as the test set. The speech data used in this work is manually transcribed at the phoneme level by experienced annotators with sufficient linguistic training. The data distribution is summarized in Table I. The correct tokens represent the phonemes pronounced correctly, and the incorrect tokens represent the phonemes pronounced incorrectly.

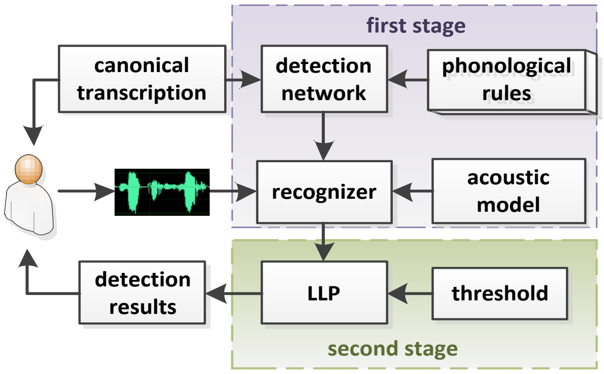

Fig. 2. The flow chart of two-stage mispronunciation detection

In our experiments, each acoustic feature is composed of 13 PLP and their first and second order derivatives, and Cepstrum Mean Normalization (CMN) is adopted. The train subset of TIMIT corpus is used to train the cross-word triphone HMMs. Each of the HMMs is tri-state model with 12 Gaussian mixtures trained with HTK.

B. Mispronunciation detection with phonological rules

Search the optimal value of α

To find the optimal criterion, three rule ranking criteria are used respectively to detect mispronunciation with ASR.

Firstly we investigate the influence of α in F_α-score and new F_α-score. The F-score measure is used to evaluate the performance of the top-N rules:

$$F=\frac{2PR}{P+R}, \tag{6}$$

where P is the precision of top-N rules, and R is the recall of top-N rules. P and R refers to the observed ones from the data.

The *F*-score of the selected top-N rules ranked by the F_α-score or trigger criterion is shown in Fig. 3. It can be seen that the curve for α =1 is next to that of ranked by trigger. Also, for α smaller than 0.01, the performance difference becomes very small. These indicate that:

- when α is large (approaching 1), although recall can be increased, it indeed degrades precision more significantly and the F-score is low as a result of that.;
- when α is too small (<0.01), it will over-emphasize precision while degrades recall more significantly, this can also make the F-score very low.

From this experiment result we empirically choose 0.01 as the optimal value of α for the following experiments.

Detect with the phonological rules ranked by the three criteria

In the second experiment we test the three different ranking criteria. The performance metric used here are false acceptance rate (FAR), false rejection rate (FRR) and total detection errors (TotErr), which are defined as follows:

$$FAR=\frac{FA}{FA+TR}, \quad FRR=\frac{FR}{FR+TA}, \tag{7}$$
$$TotErr=FA+FR$$

TABLE I DATA DISTRIBUTION OF THE TRAINING AND TEST SET.

	Number of correct tokens	Number of incorrect tokens
Training set	95153	22652
Test set	96246	23838

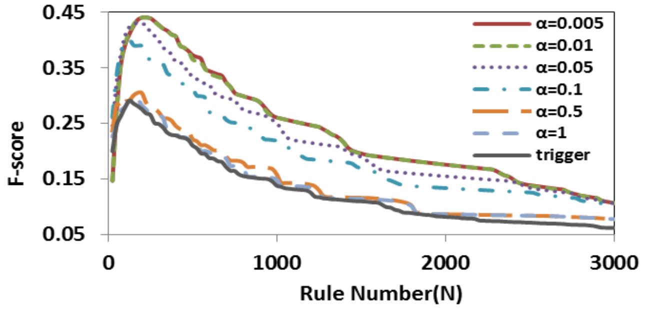

Fig. 3. F-score of top-N rules ranked by the F_α -score or trigger criterion

where FA is the number of detecting the actual mispronunciation as correct ones, FR is the number of detecting the actual correct pronunciation as errors, TA is the number of detecting the actual correct pronunciation as correct ones, and TR is the number of detecting the actual mispronunciation as errors.

The TotErr results got from the three ranking criteria are shown in Fig. 4. We can see that by every criterion, the TotErr decreases first and then keeps increasing as the number of rules increases. This is because at the beginning when the rules increase, the hits increase faster than the false alarms. But once the rules bring too much confusion to the recognition network, the false alarms will grow much faster than the hits. Overall, the new F_α -score can select the optimal rules with large hits and small false alarms. The least TotErr is obtained with the top-250 rules by the new F_α -score ranking criterion.

Comparison of different rule sets

We extract all the phonological rules from the test data, and use the basic test rule set to detect the mispronunciations on the training set, to make a comparison with the rules got from the training set. The results of different rule set are listed in Table II. The performance with basic training rule set is worse than that with testing rule set. But through rule pruning, the performance of optimal rule set (the top-250 rules by the new F_α -score ranking criterion) is greatly improved, with higher FAR and lower FRR. Because the false rejection will cause more negative effect to the learner than the false acceptance, we prefer to use the optimal rule set to reduce the FRR.

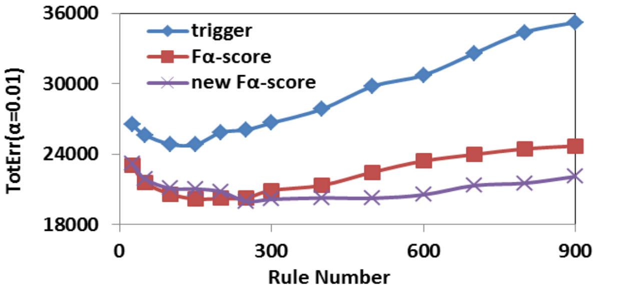

Fig. 4. The total detection errors for the three different ranking criteria

TABLE II. THE DETECTION RESULTS OF DIFFERENT RULE SET

Rule set	FAR%	FRR%	TotErr
Basic test rule set	40.11	37.04	55704
Basic training rule set	41.63	37.1	56351
Optimal rule set (new F_α -score 250 rules)	68.72	3.84	**19991**

C. *Mispronunciation detection with NLPP*

In this section, we search an optimal operating point for the NLPP. Firstly, we ignore the effect of different phoneme, and use the same threshold for all of the phones. In this condition, the equal error rate (EER) operating point is often used [10]. As our goal is to minimize the TotErr in the mispronunciation, the minimal TotErr operating point is employed.

However, in many occasions the optimal thresholds for phonemes are different, which is caused by the different acoustic characteristic of the phones. So here we also show the TotErr with and without the effect of different phonemes. The phone-dependent threshold is obtained with minimal TotErr operating point. In addition, the LPP is compared to NLPP. The detection results with different thresholds for NLPP and LPP are given in Table III. The NLPP with phone-dependent threshold is proved to be the most effective.

D. *Two-stage mispronunciation detection*

Having analyzed the performance of phonological rules and NLLP method for mispronunciation detection, here we test the performance of the overall two-stage mispronunciation prediction approach. In the first stage, the ASR with phonological rules is employed to find out the most likely pronunciations and the speech segment of each pronunciation unit. Then the second stage detection utilizes the NLPP to identify each pronunciation unit, and gives the final detection results.

We choose the three optimal rule sets by different ranking criterion to detect as the baseline: (1) top-150 rules by the trigger ranking criterion, (2) top-150 rules by the F_α -score ranking criterion, (3) top-250 rules by the new F_α -score ranking criterion. The detection results combining baseline and NLPP are given in Table IV. The two-stage architecture behaves effectively to reduce the TotErr. The minimal TotErr is acquired by new F_α -score phonological rules and NLLP, which brings the improvement of FAR from 68.72% to 63.45%, with the cost of the small increase of FRR from 3.84% to 4.62%.

V. CONCLUSIONS

This paper presents a two-stage mispronunciation detection approach for CAPT. ASR with phonological rules and NLLP

TABLE III. THE DETECTION RESULTS WITH DIFFERENT THRESHOLDS FOR NLPP/LPP

	Threshold	TotErr
NLPP	Phone-independent, EER	43504
NLPP	Phone-independent, Minimal TotErr	23969
NLPP	Phone-dependent, Minimal TotErr	**22264**
LPP	Phone-dependent, Minimal TotErr	25366

TABLE IV. THE DETECTION RESULTS OF DIFFERENT DETECTION METHOD

Detection method	FAR%	FRR%	TotErr
trigger	62.74	9.93	24828
trigger + NLLP	57.86	9.71	23457
F_α-score	71.73	3.34	20195
F_α-score + NLPP	65.1	4.42	19664
new F_α-score	68.72	3.84	19991
new F_α-score + NLLP	63.45	4.62	**19498**

are combined to improve the mispronunciation detection performance. Moreover, a new ranking criterion is used to balance the mispronunciation coverage and recognition confusion, aiming at picking up the rules work well together with ASR. In the future work, it will be considered how to train the more discriminative model for speech unit and search more effective feature to distinguish the mispronunciations and the correct ones.

REFERENCES

[1] Martha C. Pennington, “Computer-aided pronunciation pedagogy: Promise, limitations, directions,” *Computer Assisted Language Learning* , Vol. 12, No. 5, pp. 427–440,1999.

[2] M. Eskenazi, “An Overview of Spoken Language Technology for Education,” *Speech Communication*, vol. 51, pp. 832-844, October.2009.

[3] H. Meng, Y. Y. Lo, L. Wang, and W. Y. Lau, “Deriving Salient Learners’ Mispronunciations from Cross-language Phonological Comparisons,” in *Proc. of ASRU*， 2007.

[4] 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 INTERSPEECH*，2008.

[5] 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 SLaTE*, 2009.

[6] W. K. LO, S. ZHANG and H. MENG, "Automatic Derivation of Phonological Rules for Mispronunciation Detection in a Computer-Assisted Pronunciation Training System," in the *Proceedings of Interspeech*, 2010.

[7] Z. Rivlin, “A Confidence Measure for Acoustic Likelihood Scores,” *Eurospeech*, Vol. 1, pp. 523-526, 1995.

[8] S.M. Witt, and S.J.Young, “Phone-Level Pronunciation Scoring and Assessment for Interactive Language Learning,” *Speech Communication*, Vol.30, pp.95-108, 2000.

[9] H. Strik, K. Truong, F.D. Wet, and C. Cucchiarini, “Comparing Different Approaches for Automatic Pronunciation Error Detection,” *Speech Communication*, Vol. 51, pp. 845-852, 2009.

[10] H. Franco, L. Neumeyer, M. Ramos, and H. Bratt, “Automatic Detection of Phone-Level Mispronunciation for Language Learning,” *Proc. of Eurospeech*, 1999.