IMPROVED APPROACHES OF MODELING AND DETECTING ERROR PATTERNS WITH EMPIRICAL ANALYSIS FOR COMPUTER-AIDED PRONUNCIATION TRAINING

Yow-Bang Wang, Lin-Shan Lee

Graduate Institute of Electrical Engineering, National Taiwan University, Taipei, Taiwan(R.O.C.)
piscesfantasy@gmail.com, lslee@gate.sinica.edu.tw

ABSTRACT

Error pattern detection is very helpful in Computer-Aided Pronunciation Training (CAPT). This paper reports the work of modeling and detecting Error Patterns defined by language teachers based on their linguist knowledge and pedagogical experiences. We develop a model generation framework to create the Error Pattern models from existing phoneme models. We also propose a serial structure for integrating Goodness-of-Pronunciation with the Error Pattern detectors. Experimental results and analysis over different approaches for modeling and detecting Error Patterns are presented, and it is found that both the binary classification error rates and the capability of Error Pattern diagnosis can be improved effectively with the proposed approaches.

Index Terms— Computer-Aided Pronunciation Training, Mispronunciation Detection, Error Pattern, GOP

1. INTRODUCTION

Computer-Aided Pronunciation Training (CAPT) has become very important in the era of a globalized world. In particular, mispronunciation detection aims at automatically locating incorrectly pronounced acoustic segments, or even specifying the type of errors the language learner has made. Such types of error are usually referred to as Error Patterns (EPs). In general, EPs are the patterns of erroneous pronunciations frequently produced by language learners, usually caused by some articulator mechanism present in the target language but missing in the native language of the learners. To derive such EPs, some research works began with the pronunciation error rules induced by experts in literatures of second language learning [1][2], and some compared the orthographic transcription to the actual pronunciation annotated by human listeners [3].

On the other hand, the posterior probability based scores, such as Goodness-of-Pronunciation (GOP) [4], are well-known pronunciation quality measures which calculate the posterior probability that the speaker uttered a certain phoneme given the corresponding acoustic segment. The GOP-based mispronunciation detection is intrinsically different from the EP-based one. While EP-based detectors choose among multiple EPs, GOP-based ones often make binary decision on whether the segment is correctly or incorrectly pronounced. Such difference also implies these two kinds of mispronunciation detectors may be complementary.

In this paper, we focus on the detection of phoneme-level EPs of Mandarin Chinese. We use model adaptation techniques to create EP models based on existing phoneme models, and then construct an efficient EP-based mispronunciation detector with these models. We further propose to integrate EP-based and GOP-based mispronunciation detectors in a serial structure, to achieve better performance than previously proposed approaches [5], by better utilizing the complementarities between EP-based and GOP-based detectors. The rest of this paper is organized as follows. Sections 2 and 3 describe our corpus and EP labeling system respectively. Section 4 explains how the EP models were generated, and Section 5 the proposed framework for mispronunciation detection. Experimental results are reported and discussed in Section 6. Concluding remarks are finally made in the last section.

2. DATA COLLECTION

We have been collaborating with Chinese language teachers from International Chinese Language Program (ICLP) of National Taiwan University (NTU), working on Computer-Aided Pronunciation Training for learning Mandarin Chinese. The corpus used in this paper was collected in year 2008 and 2009. A total of 278 ICLP learners from 36 different countries with balanced gender and a wide variety of native languages joined our recording tasks. Each learner was asked to produce 30 sentences, each containing 6 to 24 characters. The recording text prompts were chosen to cover as many Chinese syllables and tone patterns as possible, and were selected from the learning materials designed by ICLP language teachers and used in NTU Chinese [6], a successfully operating online Chinese pronunciation learning software.

We took the recordings of 186 learners as our adaptation set, 50 learners as development set, and 42 learners as testing set. In Table 1 we can see that the percentage of mispronounced segments in each data set is relatively small. This implies the learners already had some basic training of Mandarin Chinese, and lead to difficulties in this work.

Table 1. *The percentage of pronunciation errors of each data set.*

Data set	Percentage of pronunciation errors
Adaptation	10.32%
Development	8.55%
Testing	9.50%

3. ERROR PATTERN DEFINITION AND LABELING

The EPs we used was summarized by the language teachers based on their linguistic knowledge and pedagogical experiences, to cover most frequent EPs made by learners of Mandarin Chinese, and is not limited to any specific corpus. The basic unit used in our EP definition is Mandarin phonemes represented in Zhuyin. We have a total of 39 canonical Mandarin phoneme units. Tables 2 and 3 are two examples of our EP definition for consonant and vowel respectively. In the "ID" column, canonical pronunciations are coded "000"; the codes except "000" and "099" are used to denote EPs of each phoneme; and the code "099" means "none of the above", i.e. the acoustic segment is neither pronounced correctly, nor can be categorized into any of the EPs. In the "description" column, "CH" and "EN" respectively indicate Chinese or English phonemes used to describe the EP. One can see that our definition of EPs not only include phoneme-level substitution (e.g. l_010 in Table 2), but also insertion (e.g. ei_010 in Table 3) and deletion (e.g. ei_020 in Table 3).

A total of 152 EPs including "099" were defined through the discussion of a group of language teachers, and the surface pronunciation of each acoustic segment in learners' recordings was labeled as one of the patterns. Most of these EPs are described primarily using phonemes in Mandarin Chinese or English, and in some other cases using phonemes in Chinese dialects common in Taiwan (Min or Hakka languages). Note that although these EPs are described with phonemes from a certain language, these phonemes are not for specifying the L1 of learners. Learners whose native languages are not English may still utter the EPs defined with English phonemes.

Table 2. *The definition of EPs of the consonant /l/*

ID	Discription
l_000	Canonical pronunciation.
l_010	Pronounced as EN_l.
l_020	Pronounced as CH_r.
...	...
l_099	None of the above.

4. ERROR PATTERN MODEL GENERATION

To obtain the acoustic models of these EPs, one intuitive approach would be to use existing phoneme models based on their descriptions. However, we have found that the resulting performance of such intuitive approach was very poor. This may be because the description of an EP is just the closest representation of its acoustic realization, but not its exact surface pronunciation; and the mismatch among different corpora is also a problem.

Table 3. *The definition of EPs of the diphthong /ei/*

ID	Discription
ei_000	Canonical pronunciation.
ei_010	Pronounced as CH_u+CH_ei.
ei_020	Pronounced as CH_e (lack of the ending CH_i).
...	...
ei_099	None of the above.

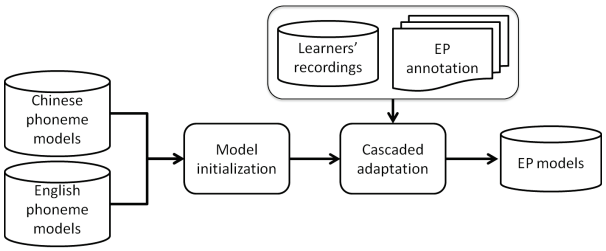

Fig. 1. *Block diagram of EP model generation in the proposed approach.*

Fig. 1 illustrates how we generated EP models out of the phoneme models we have at hand. First we initialized each EP model with either of these two strategies below:

1. Homogeneous initialization: Each canonical pronunciation model is duplicated as the initial models of its corresponding EPs.

2. Heterogeneous initialization: The Chinese or English phoneme models of those phonemes used in the description of the EPs are copied and renamed as the initial EP models. If the EP is described with Min or Hakka phonemes, since we do not have models for Min or Hakka phonemes, we pick the most similar Chinese phoneme models as the initial EP models based on the suggestions from language teachers.

Then cascaded adaptation [7], which includes three stages: global Maximum Likelihood Linear Regression (MLLR), class-based MLLR and Maximum A Posteriori (MAP) adaptation cascaded in sequence, was performed with the adaptation data set after model initialization. The copies of existing phoneme models are therefore transformed into EP models.

5. MISPRONUNCIATION DETECTION

5.1. EP-based detector using pronunciation network

Once we have the acoustic models of the EPs, phone-level forced alignment can be performed with learners' recordings. We expanded the transcribed phoneme sequence into a network of canonical pronunciations and EPs. The surface

pronunciations with maximum likelihood are automatically chosen during forced alignment. As mentioned in Section 3, our EPs not only include phoneme-level substitution, but phoneme-level insertion and deletion as well. We also inserted a short-pause (sp) model between every two syllables to capture the possible hesitation the learners may have.

5.2. GOP-based detector with pre-defined threshold

The GOP defined in this paper is:

$$GOP(O^{(p)}) = \frac{1}{D_p} log P(p|O^{(p)}) \tag{1a}$$

$$\approx \frac{1}{D_p} log \left(\frac{P(O^{(p)}|p)}{\sum_{q \in Q} P(O^{(p)}|q)} \right) \tag{1b}$$

$$\approx \frac{1}{D_p} log \left(\frac{P(O^{(p)}|p)}{max_{q \in Q} P(O^{(p)}|q)} \right), \tag{1c}$$

where $O^{(p)}$ is the acoustic segment of phone p, D_p is the duration of the segment, and q is a phoneme model out of the set of all phonemes Q. The numerator of GOP is derived using forced alignment, and the denominator is derived using free-phone recognition with time boundaries constrained by forced alignment result. In our experiments, the phoneme model set Q was the canonical pronunciation models adapted to the correctly pronounced segments in the adaptation set.

GOP can be used in mispronunciation detection with a pre-defined threshold. All the segments with scores lower than the threshold are classified as mispronunciations. The threshold can be single-valued and equally applied to all phonemes, or can be phoneme-dependent.

5.3. Integrating GOP-based and EP-based detectors

Fig. 2 (a) illustrates the serial structure for integrating EP-based and GOP-based detection proposed here. The acoustic segment is first evaluated by the EP-based detector. If EP-based detector accept it as correct pronunciation, this segment is further passed to GOP-based detector for double-check. A segment is predicted as correct pronunciation only if it is accepted by both detectors; otherwise, this segment is classified as the EP with maximal likelihood except the canonical pronunciation. Also, the parallel structure proposed previously is illustrated in Fig. 2 (b), in which a list of phonemes better handled by GOP-based detector is first obtained, and the acoustic segments of such phonemes are then backed-off without EP diagnoses , i.e. each acoustic segment is checked by a single detector in testing phase despite both detectors are available [5].

One advantage of the proposed serial structure is the phoneme-dependent GOP thresholds can be tuned to optimize the performance not only for the GOP-based detector itself, but for the integrated system. In this way GOP is used as a compensation stage for EP-based detector. Another advantage is that EP diagnostic feedback is always available to the learners, no matter the segment is rejected by EP-based or GOP-based detector; on the other hand, the parallel structure could offer EP diagnosis only when the phoneme is better evaluated by the EP-based detector.

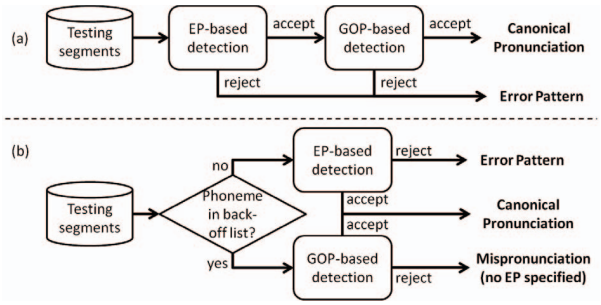

Fig. 2. *(a) Serial structure and (b) parallel structure for integrating EP-based and GOP-based mispronunciation detection.*

6. EXPERIMENTS

6.1. Acoustic model training corpus

Our Chinese phoneme models were trained using the ASTMIC Mandarin corpus of read speech recorded by 95 males and 95 females, each reading 200 sentences, with a total length of 24.6 hours; and our English phoneme models were trained using the training set of TIMIT corpus recorded by 462 speakers from eight dialect regions of the USA. We chose monophone as our acoustic model unit. Most Zhuyins are based on monophones except some diphthongs. For diphthongs we modify our lexicon so that a diphthong can be mapped to two or more consecutive monophone models.

6.2. Performance measure

The False Rejection Rate (FRR) and False Acceptance Rate (FAR) are defined as below:

$$FRR = \frac{FalseRejection}{FalseRejection + TrueAcceptance}, \tag{2a}$$

$$FAR = \frac{FalseAcceptance}{FalseAcceptance + TrueRejection}, \tag{2b}$$

and the Average Error Rate (AER) is further calculated:

$$AER = \frac{FAR + FRR}{2}. \tag{3}$$

In addition to these binary classification error rates, the number of instances which are correctly diagnosed (CD) is also collected.

6.3. Experimental results and discussion

Table 4 lists the results of GOP-based (row (1)), EP-based (row (2)(3)) and the two integration structures (row (4)(5)).

The EP-based$_{homo}$ and EP-based$_{hetero}$ stand for EP-based detectors constructed with homogeneous and heterogeneous initialization respectively. For the GOP-based detector and both integration structures, phoneme-dependent GOP thresholds were tuned to minimize the AER, and the best operating points are reported here. We further illustrate the relation between FAR and FRR of these approaches in Fig. 3. The extra curve in Fig. 3 is served as a baseline of our system, which is the GOP-based detector with single-valued threshold applied equally to all phonemes, and different points on this curve represent different values of the threshold.

Comparing two EP-based detectors in Table 4, we can see that both FAR and FRR of EP-based$_{hetero}$ are lower, and the number of correct diagnosis (CD) given by EP-based$_{hetero}$ is also larger. This may due to the EP models started with heterogeneous initialization were intrinsically different, and thus resulted in better discriminability. Also, compared to the GOP-based detector, both the EP-based detectors yielded obviously lower FRR but higher FAR. This implies our EP-based detectors are prone to accept testing segments as canonical pronunciation, even if they actually belong to some EPs. This is obviously because the percentage of mispronounced segments in our corpus is relatively small, and the EP models obtained therefore suffered more from the mismatch among data sets than canonical pronunciation models.

Now consider the two integration approaches. We used EP-based$_{hetero}$ as the EP-based detector to be integrated. With parallel structure, the AER became worse compared to EP-based$_{hetero}$, and the CD was also decreased because some phonemes were backed-off without EP diagnosis; On the other hand, the proposed serial structure further reduced the AER, and the CD is also slightly increased. Note that the FAR is largely reduced with the serial structure. This shows how the weakness of EP-based detector can be successfully compensated by GOP.

Table 4. *The experimental results of mispronunciation detection of our testing set.*

	FAR	FRR	AER	CD
(1) GOP-based	42.43%	23.60%	33.01%	-
(2) EP-based$_{homo}$	47.88%	14.05%	30.96%	1108
(3) EP-based$_{hetero}$	47.06%	12.58%	29.82%	1126
(4) EP+GOP, parallel	42.15%	18.77%	30.46%	948
(5) EP+GOP, serial	34.35%	23.65%	29.00%	1134

7. CONCLUSION

In this paper we introduced our framework of modeling and detecting EPs with empirical analysis over different modeling and detecting approaches. Using different model initialization strategies and the cascaded adaptation, we generated EP models from existing phoneme models. We also investigated different structures for integrating EP-based and GOP-based mispronunciation detectors. Extensive experimental results showed the EP models initialized heterogeneously based on phoneme models from different languages are more discriminative, and the proposed serial structure for integrating EP-based and GOP-based mispronunciation detectors further reduced the overall binary classification error rates and improved the capability of EP diagnosis.

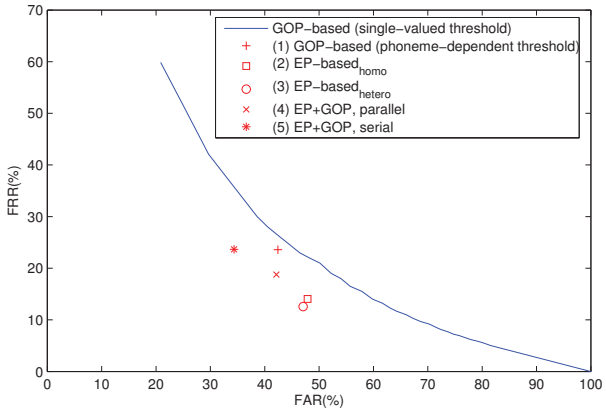

Fig. 3. *The FAR and FRR of different approaches of mispronunciation detection.*

8. REFERENCES

[1] A. Harrison, W. Lo, X. Qian, and H. Meng, "Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training," in *Sig-SLATE*, 2009.

[2] C. Cucchiarini, H. Van Den Heuvel, E. Sanders, and H. Strik, "Error selecion for asr-based english pronunciation training in 'my pronunciation coach'," in *Proc. INTERSPEECH 2011*.

[3] D. Luo, X. Yang, and L. Wang, "Improvement of segmental mispronunciation detection with prior knowledge extracted form large l2 speech corpus," in *Proc. INTERSPEECH 2011*.

[4] S. Witt and S. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, no. 2-3, pp. 95–108, 2000.

[5] H. Meng, W. Lo, A. Harrison, P. Lee, K. Wong, W. Leung, and F. Meng, "Development of automatic speech recognition and synthesis technologies to support chinese learners of english: The cuhk experience," in *APSIPA Annual Summit and Conference 2011*.

[6] NTU Chinese. [Online]. Available: http://chinese.ntu.edu.tw/

[7] C. Yeh, C. Huang, and L. Lee, "Bilingual acoustic model adaptation by unit merging on different levels and cross-level integration," in *Proc. INTERSPEECH 2011*.