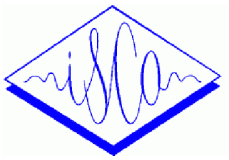
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Error Pattern Detection Integrating Generative and Discriminative Learning 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

Computer-Assisted Language Learning tries to have computers serve as virtual language tutors to help people in learning non-native languages in the globalized world nowadays. In this paper we propose a framework to incorporate specially designed discriminative models with carefully trained generative models for the task of pronunciation error pattern detection. For each phoneme we train one or more SVMs with varying targets and different weights to integrate with HMM/GMMs for optimizing the detection performance from different aspects. Experiments show this integration framework effectively enhance mispronunciation detection performance.

Index Terms: Error Pattern, Viterbi Decoding, Discriminative Score, SVM

10.21437/Interspeech.2012-249

1. Introduction

Computer-Aided Pronunciation Training (CAPT) has been extensively studied because speech processing technology can be very useful in training the pronunciation of language learners. Lots of CAPT systems have been available for such purposes. In these systems, in order to generate useful feedbacks for the learners to improve their language skills, reliable mispronunciation detection is crucial. Mispronunciation detection aims at automatically locating mispronounced acoustic segments, or even specifying the way the mispronunciation was made, for example in terms of 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. One popular approach of EP detection is to consider EPs as variants of canonical pronunciation. By expanding the transcribed phoneme sequence of the produced utterance into a network of canonical pronunciations and EPs, the surface pronunciations with maximum likelihood are automatically chosen using phoneme-level forced alignment.

On the other hand, EP detection can also be considered as a classification problem. Given the boundaries of the phoneme segments in learners' utterances, the task is actually to determine whether each segment is correctly pronounced or not, or determine which EP each segment belongs to. Log-likelihood ratio or posterior-probability based scores such as Goodness-of-Pronunciation (GOP) are well-known measures of pronunciation quality for such purpose. Some works used improved GOP with predefined thresholds to find mispronounced segments [1], and some further incorporated GOP-based mispronunciation detector with EP network to boost the performance [2][3].

The difficulty of EP detection comes from the fact that EPs are intrinsically similar to their corresponding canonical pronunciation, and to distinguish between different EPs of the same phoneme can be even harder than to distinguish between different phonemes. In searching for better discriminability, more powerful classifiers such as Support Vector Machine (SVM) have been widely used in mispronunciation detection, often with log-likelihood ratio or posterior probability vectors as input features [4][5][6]. Other applications of SVM in the area of pronunciation evaluation included using SVM-Rank to quantify multiple levels of proficiency [7], or using the output posteriorgram of SVM as features concatenated with other acoustic features to estimate the pronunciation score and intelligibility [8].

In this paper, we propose a framework to closely integrate generative models (Hidden Markov Model/Gaussian Mixture Model (HMM/GMM) for each EP [3]) and discriminative models (SVM with Radial Basis Function (RBF)) for EP detection. The rest of this paper is organized as follows. Section 2 presents the architecture and formulation of the proposed framework. Section 3 introduces the corpus used and training procedure. Experimental results are reported and discussed in Section 4. Concluding remarks and some future directions are finally noted in the last section.

2. Proposed Approach

2.1. Baseline: generative EP modeling

We have been working on developing CAPT software for learners to practice their Mandarin pronunciation. This work is part of a joint project with the International Chinese Language Program of National Taiwan University. The corpus used here includes recordings from learners in this program, and the EPs are manually labeled by the language teachers. Most of the EPs are defined in terms of Mandarin and English phonemes.

We have proposed recently a special approach to explicitly derive the HMM/GMM acoustic model of each

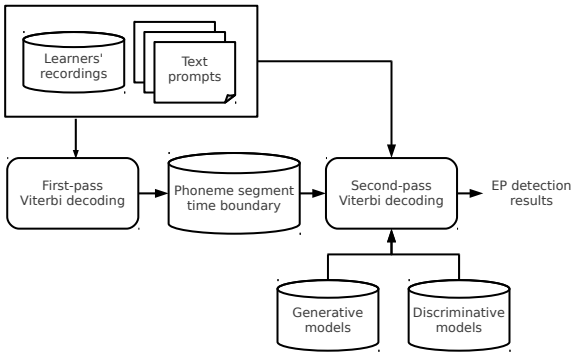


Figure 1: *The proposed framework for integrating generative and discriminative models for EP detection.*

EP [3]. Based on the definition of each EP given by language teachers, we duplicated corresponding Mandarin or English phoneme models trained from some other native Mandarin and English corpus as the initial EP models. Such model initialization ensures that the EP models are intrinsically different from each other, and thus yields better discriminability. These initial EP models are then adapted by the EP segments in the learners' corpus and transformed into final EP models. We use these EP models to construct the pronunciation network for maximal-likelihood alignment, and the surface pronunciation of learners' recordings can thus be determined. This is the baseline system in the experiments reported here.

2.2. Proposed framework of integrating generative and discriminative models

Fig. 1 illustrates the proposed framework for integrating generative and discriminative models for EP detection. First we perform first-pass Viterbi decoding to obtain the time boundaries for the phoneme segments in learners' utterances. This first-pass forced alignment can be performed using either the ordinary acousic model set, or the EP model set as described in Sec.2.1. In our experiments we used the EP model set for first-pass forced alignment to get more precise time boundaries. Then the second-pass Viterbi decoding is performed with known segment boundaries for EP detection.

We incorporate both generative and discriminative models in the second-pass Viterbi decoding. Let e_p^i be the i^{th} EP for a phoneme p, $i = 0, 1, 2, ..., N_p$, where e_p^0 is the canonical pronunciation and N_p is the total number of EPs for phoneme p. For each frame x of 39-dimensional MFCC vector, the acoustic score $S(x, e_p^i)$ of x w.r.t. e_p^i is given by:

$$S(x, e_p^i) = S_g(x|e_p^i) + S_d(e_p^i|x), \tag{1}$$

where $S_g(x|e_p^i)$ is the log-likelihood score of e_p^i given by generative model (HMM/GMM), and $S_d(e_p^i|x)$ is the score of e_p^i given by discriminative model (SVM). The HMM/GMMs are exactly the same as in the baseline approach as described in Sec.2.1, and the detail of discriminative score is explained below.

2.3. Discriminative modeling and scoring with SVM

We propose three different schemes for discriminative modeling and scoring using SVMs:

(a) (N_p+1)-ary SVM: For each phoneme p , an (N_p+1)-ary SVM is trained to classify each frame into either the canonical pronunciation e_p^0 or one of the EPs $e_p^i, i = 1, 2, ..., N_p$. The discriminative score used in Eq.(1) is then simply the weighted log-posterior of the (N_p+1)-ary SVM:

$$S_d(e_p^i|x) = w_p \cdot ln(P(e_p^i|x)). \tag{2}$$

Note here a distinct SVM is trained for each phoneme. We allowed the weights w_p to be 0 in case some SVMs are ill-trained.

(b) Binary SVM: For each phoneme p, a binary SVM is trained to classify each frame into either correct or incorrect pronunciation, and

$$S_d(e_p^i|x) = \begin{cases} w_p \cdot ln(P(correct|x)), & \text{for i=0,} \\ w_p \cdot ln(P(incorrect|x)), & \text{otherwise.} \end{cases} \tag{3}$$

Note that this scheme is in favor of rejecting testing segments as mispronunciation. Because

$$P(incorrect|x) = \sum_{i=1}^{N_p} P(e_p^i|x), \tag{4}$$

$P(incorrect|x)$ is always an overestimation of each individual EP posterior. Such design is based on the observation in our previous work that generative model based detectors are usually prone to accept testing segments as canonical pronunciation, even if they actually belong to some EPs [3]. Therefore this discriminative score is introduced to leverage the inclination toward false acceptance.

(c) Cascaded SVMs: For each phoneme p, two SVMs are trained and cascaded. The first is binary for classifying each frame into either correct or incorrect pronunciation, and the second is N_p-ary for classifying each frame into one of the EPs except the correct pronunciation. The discriminative score is given by:

$$S_d(e_p^i|x) = \begin{cases} w_{1,p} \cdot ln(P_1(correct|x)) & \text{, if i=0,} \\ w_{1,p} \cdot ln(P_1(incorrect|x)) & \\ +w_{2,p} \cdot ln(P_2(e_p^i|x)) & \text{, otherwise,} \end{cases} \tag{5}$$

where $P_1(\cdot)$ is from the first binary SVM and $P_2(\cdot)$ from the second N_p-ary SVM. This scheme is an extension from the previous one, aiming at more discriminative estimation for the scores of EPs.

3. Corpus and Implementation

3.1. EP definition and data collection

The acoustic units for EP definition in this work are Mandarin phonemes represented in Zhuyin. There is a total of 39 canonical Mandarin phoneme units, and 152 EPs were summarized by language teachers based on their linguistic knowledge and pedagogical experiences, to cover most frequent EPs made by Mandarin Chinese learners. The definition of EPs not only include phoneme-level substitution, but also insertion and deletion, and is not limited to any specific corpus [3].

The Mandarin Chinese learners' corpus used in this work was collected in year 2008 and 2009. 278 learners studying Mandarin Chinese from 36 different countries with balanced gender and a wide variety of native languages joined the recording tasks. Each learner was asked to produce a set of 30 phonetically balanced and prosodically rich sentences, each containing 6 to 24 characters. These 30 sentences covered almost all frequently used Mandarin syllables and tone patterns.

We took the recordings of 186 learners with a total length of 8.4 hours as the set for both SVM training and acoustic model adaptation, 50 learners with 2.2 hours as the development set, and 42 learners with 2.0 hours for testing. The surface pronunciation of each acoustic segment in learners' recordings was labeled by the language teachers as correct pronunciation or one of the EPs. The percentage of mispronounced segments in each data set is relatively small [3]. This implies most learners here already had some basic training of speaking Mandarin Chinese, which led to the limited quantity of data for inferencing reliable EP models in this work.

3.2. Performance measure

To equally emphasize the performance of the proposed approach for each phoneme, we define the Average False Rejection Rate (AFRR) and Average False Acceptance Rate (AFAR) as the average of False Rejection Rate ($FRR(p_i)$) and False Acceptance Rate ($FAR(p_i)$) for each phoneme p_i across all different phonemes $p_1, ..., p_M$:

$$AFRR = \frac{1}{M}\sum_{i=1}^{M} FRR(p_i), \tag{6a}$$

$$AFAR = \frac{1}{M}\sum_{i=1}^{M} FAR(p_i). \tag{6b}$$

The Average Error Rate (AER) is then calculated as:

$$AER = \frac{AFAR + AFRR}{2}. \tag{7}$$

In addition to the above binary classification error rates, the numbers of mispronunciation instances which are correctly diagnosed (CD) as the transcribed EP or incorrectly diagnosed (DE) as different EP are also collected.

3.3. EP HMM/GMM training

The Mandarin phoneme models for EP initialization were trained with the ASTMIC Mandarin corpus of native read speech produced by 95 males and 95 females, each with 200 utterances, with a total length of 24.6 hours; while the English phoneme models for EP initialization were trained using the training set of TIMIT corpus produced 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 modified our lexicon so that a diphthong can be mapped to two or more consecutive monophone models.

With these Chinese and English phoneme models, and the adaptation set as described in Sec.3.1, the HMM/GMMs for EPs were derived as explained in Sec.2.1. We use these EP models in our baseline EP detection system, and in the first-pass and second-pass Viterbi decoding in the proposed framework as shown in Fig.1.

3.4. SVM training and weight tuning

In Sec.2.3, there are two elements to be determined: the required SVMs and the corresponding weights. In our experments, the training set of SVMs was the same as the adaptation set for EP HMM/GMMs. For SVM training, the input instances are 39-dimensional MFCC vectors, and the classification target is varied as described in Sec.2.3. We use libSVM for the implementation of SVM, and chose RBF kernel for its well-known discriminative power.

One problem arose when training the (N_p+1)-ary SVM and binary SVM: the number of correctly-pronounced instances are far more than the instances of EPs. To alleviate this data imbalance problem, we down-sampled the correctly-pronounced instances to the number of second-largest class, which is the mispronunciation class for binary SVM, and the EP with the most instances for (N_p+1)-ary SVM.

The parameters of SVMs (σ and C for RBF kernel) were tuned based on the development set to minimize frame AER for (N_p+1)-ary and binary SVMs, and minimize the ratio between DE and CD for N_p-ary SVMs. After the optimal SVMs of each phoneme were determined, we embedded the SVMs into the proposed architecture as in Fig.1, and the weights of SVMs were then tuned to minimize segment AER of the overall system on the development set.

4. Experimental Results

Table 1 shows the phoneme segment level binary classification results of our baseline and three proposed schemes on the testing set, and table 2 the EP diagnosis results. Overall speaking, the AFAR is effectively reduced with all three proposed schemes, and the reduction was significant especially for scheme (c) with cascaded SVMs. However, the AFRR is inevitably increased to some extent due to the trade-off relation between false acceptance and false rejection. But in any case the average performance (AER) is reasonably improved in all schemes.

Comparing the 3 different proposed schemes, scheme (a) using an (N_p+1)-ary SVM for each phoneme resulted in the worst AER, which is only slightly better than the baseline, and the number of instances of correct diagnosis (CD) and diagnostic error (DE) are both worse than the baseline. This is definitely due to the data imbalance problem, i.e. much more data for the canonical pronunciation but too few for each EP, which made it difficult to accurately estimate the posterior probability of the canonical pronunciation and the EPs at the same time. On the other hand, scheme (b) using binary SVM resulted in obviously lower AER. This shows the powerfulness of discriminative models in effectively enhancing the performance. Yet the CD and DE are both increased compared to the baseline, which means some of the instances recovered from false acceptance were still incorrectly diagnosed, apparently because the discriminative models here are only for correct/incorrect pronunciation classification.

With an extra N_p-ary SVM for each phoneme, we expected scheme (c) could better diagnoses the EPs. Unfortunately, it turns out that CD was increased while DE was increased too. This may be because although some instances originally falsely accepted were now found incorrect, but their exact EPs were really hard to detect in the first place. Nevertheless, we did achieve the lowest AER with this scheme (c). In other words, despite the additional N_p-ary SVM failed in better discriminating among different EPs, the quality of Viterbi decoding was actually improved with finer estimation of EP posteriors.

Table 1: *The experimental results for mispronunciation detection.*

Integrated SVM	AFAR	AFRR	AER
None (baseline)	60.70%	12.03%	36.37%
(a) (N_p+1)-ary	56.65%	15.93%	36.29%
(b) Binary	47.81%	21.75%	34.78%
(c) Cascaded	41.52%	25.34%	33.43%

5. Conclusion and Future Work

In this paper, we integrate the scores from both generative and discriminative models with a two-pass Viterbi decoding architecture for better EP detection in CAPT. We designd 3 different schemes of using SVMs in the Viterbi decoding to emphasize different aspects of EP detection, while maintaining the flexibility for fine tuning. The experimental results showed that integrating SVMs with HMM/GMMs can effectively improve the EP detection. Some future directions of this work include trying to apply discriminative training [9] to obtain better EP HMM/GMMs, or further incorporating GOP-based detection [3] into this framework.

Table 2: *The experimental results of EP diagnosis.*

Integrated SVM	CD	DE
None (baseline)	1126	506
(a) (N_p+1)-ary	1100	524
(b) Binary	1217	604
(c) Cascaded	1330	683

6. References

[1] F. Zhang, C. Huang, F. Soong, M. Chu, and R. Wang, "Automatic mispronunciation detection for mandarin," in *Proc. ICASSP 2008*, pp. 5077–5080.

[2] 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*.

[3] Y. Wang and L. Lee, "Improved approaches of modeling and detecting error patterns with empirical analysis for computer-aided pronunciation training," in *Proc. ICASSP 2012*.

[4] J. Jiang and B. Xu, "Exploring the automatic mispronunciation detection of confusable phones for mandarin," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 4833–4836.

[5] Y. Chen, C. Huang, and F. Soong, "Improving mispronunciation detection using machine learning," in *Proc. ICASSP 2009*, pp. 4865–4868.

[6] S. Wei, G. Hu, Y. Hu, and R. Wang, "A new method for mispronunciation detection using support vector machine based on pronunciation space models," *Speech Communication*, vol. 51, no. 10, pp. 896–905, 2009.

[7] L. Chen and J. Jang, "Automatic pronunciation scoring using learning to rank and DP-based score segmentation," in *Proc. INTERSPEECH 2010*.

[8] H. Kibishi and S. Nakagawa, "New feature parameters for pronunciation evaluation in english presentations at international conferences," in *Proc. INTERSPEECH 2011*, pp. 1149–1152.

[9] X. Qian, F. Soong, and H. Meng, "Discriminative acoustic model for improving mispronunciation detection and diagnosis in computer-aided pronunciation training (CAPT)," in *Proc. INTERSPEECH 2010*.