EXPLORING THE AUTOMATIC MISPRONUNCIATION DETECTION OF CONFUSABLE PHONES FOR MANDARIN

Jie Jiang, Bo Xu

Institute of Automation, Chinese Academy of Sciences, Beijing, China
{jjiang, xubo}@hitic.ia.ac.cn

ABSTRACT

Mispronunciation detection is one of the vital tasks of the CALL (Computer Assisted Language Learning) systems. Many methods have been introduced to accomplish this task. However, few of them have addressed the detection task on confusable phones. In this paper, phone-level classifiers are utilized to improve the detection performance on the confusable phones. Features of the classifiers are posterior probability vectors calculated from their corresponding acoustic models. Moreover, confusion matrix is also extracted and incorporated to calculate derivatives of the posterior probability vectors. Experiments on our Mandarin database validate the effectiveness of our proposed method, compared with the commonly used posterior probability and phone dependent thresholds methods.

Index Terms— Computer Assisted Language Learning (CALL), automatic mispronunciation detection, enhanced posterior probability vector, confusion matrix

1. INTRODUCTION

In the last decade, many researchers have addressed the implementation of Computer Assisted Language Learning (CALL) systems, which aim to facilitate the language learners in the whole learning process [1]. As a virtual teacher, most of the CALL systems aim to evaluate the learners' language abilities according to their speeches. However, in certain interactive learning tasks, such as the pronunciation learning, feedbacks are even more important than proficiency scores. To improve learners' pronunciation, correctness of each uttered phone is more informative than overall scores. Under these circumstances, the goal of mispronunciation detection is to automatically label each uttered phones by correct or incorrect.

Following the GOP (Goodness of Pronunciation) score used by Witt [1], many posterior probability based methods have been investigated. Franco [2] proposed to use posterior-based methods with native models rather than log-likelihood ratio based methods with nonnative models in detection tasks. In [3], Ito utilized decision tree based error clustering and adopted multi-threshold for clusters. Zheng proposed weighted phone posterior score and showed its effectiveness on utterances with 2-3 syllables [4]. To get a better statistical model, speaker adaptive training and selective maximum likelihood linear regression were incorporated and proved to be effective by Zhang [5].

Most of the above methods have got promising results. However, few of them have paid attention to the detection of confusable phones, which are undistinguishable for the commonly used posterior probability methods. Details of the confusable phones are presented in Section 4.1.

In this paper, phone-level classifiers are introduced to improve the performance on confusable phones. Features of the classifiers are posterior probability vectors collected from their corresponding acoustic models. Further more, the features are enhanced with confusion matrix extracted from speech corpus.

In Mandarin, it's worth noting that each of the phones belongs to either of the following phone classes: the initials or the finals. As is stated by Zhang [5], each Mandarin syllable normally consists of three parts: an initial, a final, and a tone. However, this paper is focused on the phones, while the tones are ignored.

The rest of the paper is organized as follows: As a baseline, Section 2 first discusses the commonly used posterior probability based methods. And then our proposed method is explained in detail in Section 3. In Section 4, our database is presented along with the confusable phones, and the performance evaluation is also examined. Finally, a further discussion is given in Section 5.

2. BASELINE

Before detailing into our method, commonly used posterior probability based methods are examined in this Section.

2.1. Posterior Probability (PP)

In both pronunciation evaluation and mispronunciation detection, the log form of posterior probability is commonly accepted as an effective indicator of the pronunciation quality. In this paper, it is calculated as follows. Firstly, speeches and transcriptions are aligned by the Viterbi search. Secondly, for each alignment, acoustic scores of their

corresponding phone models are calculated respectively. Finally, the posterior probability (PP) is given by:

$$PP(p \mid o_s^t) = \log(\frac{P(o_s^t \mid p)P(p)}{\sum_{q \in C(p)} P(o_s^t \mid q)P(q)}) \quad (1)$$

where p is the hypothesized phone in the transcriptions, o_s^t is its observation, *C(p)* is its competitive phone set, and *P(q)* is the prior probability of the phone q. Practically, PP ought to be normalized by the frame duration $Dur(o_s^t)$.

Given the PP, correctness of each phone is determined by a global threshold, which is yielded from the distribution of the PP on the training set. The hypothesized phone would be classified as correct only if its PP is no less than the global threshold, or incorrect otherwise.

During the forced alignment, robust phone boundaries can be achieved by extending paths according to the phone class of the current phone. And the competitive phone set *C(p)* of phone p is set to its phone class, which would either be the initials or the finals.

2.2. Phone Dependent Thresholds (PDT)

As is indicated by Witt [1], phones tend to have different dynamic ranges of their PP. So a better performance could be achieved if phone dependent thresholds (PDT) are adopted. However, the requirement on a large database with evenly distributed phones usually makes it impractical in use. Nevertheless, a commonly adopted approach is to back off those thresholds of minor phones to a global one.

3. PHONE-LEVEL CLASSIFIERS

By extending the notion of phone dependent thresholds, our proposed method is building a classifier for each phone to determine their correctness.

3.1. Overview

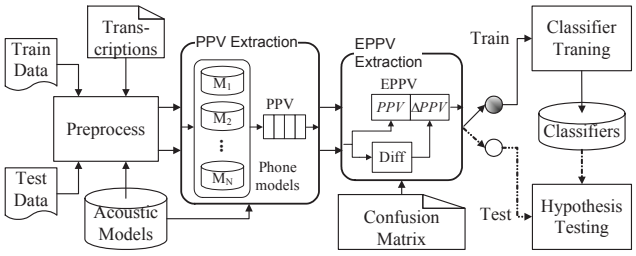

Figure 1. The proposed features and classifiers

Overview of our proposed method is illustrated in Figure 1. First, training and testing data are aligned with transcriptions in the preprocess module. Sequentially, PPV (Posterior Probability Vector defined in Section 3.2) features are extracted from the aligned observations. After that, EPPV (Enhanced Posterior Probability Vector defined in Section 3.3) features are extracted from the PPV with confusion matrix. At last, the features of train and test data are fed into classifiers for training and testing respectively. Note that both PPV and EPPV features are alternative inputs for the classifiers, and their performances are discussed in Section 4 respectively.

3.2. Posterior Probability Vector (PPV)

Extraction process of the PPV is presented as follows. Suppose there are N phones, for each phone p_i, its alphabetically sorted competitive phone set is denoted by vector D_i, $1 \le i \le N$. Suppose M_p as the acoustic model for phone p, given segment o_s^t correspondent with the phone p_j, the posterior probability vector (PPV) could be calculated as:

$$\begin{aligned} PPV(p_j \mid o_s^t) &= [v_1 \quad v_2 \quad \dots \quad v_K]^T \\ v_i &= F_{D_j^i}(p_j \mid o_s^t) \end{aligned} \quad (2)$$

$$F_{D_j^i}(p_j \mid o_s^t) = \frac{1}{Dur(o_s^t)} \log(\frac{P(o_s^t \mid M_{p_j})P(p_j)}{\sum_{q \in C(D_j^i)} P(o_s^t \mid M_q)P(q)}) \quad (3)$$

where K is the size of D_j, and D_j^i is the i-th element of D_j. Size of the PPV is determined by the target phone p_j. Thus, mispronunciation detection can be extended into a vector space of dimension K for each phone.

Note that in formula (3), $F_{D_j^i}(p_j \mid o_s^t)$ will equal to the normalized $PP(p \mid o_s^t)$ if D_j^i equals to j. It means that the PP is one of the elements of the PPV, and information in the PP is equally carried by the PPV. Consequently, the methods presented in Section 2 are simplifications of the phone-level classifiers. With the PPV, detection procedure is expanded from single feature to feature spaces.

3.3. Enhanced Posterior Probability Vector (EPPV)

Performance of the PPV can be enhanced by adding spatial derivatives to reflect the varieties of PP within models. However, suppose K is the size of the PPV, appending derivatives of its elements increases the size by C_K^2. To decrease the impact of data sparseness, the most similar phones should be selected to calculate the derivatives.

3.3.1 Confusion Matrix

Confusion knowledge has been used in ASR systems to represent confusion rules. In [6], it was extracted using the two time-aligned transcriptions given by the spoken and by the native language ASRs. However, in this paper, the process is modified to create a matrix. Firstly, each utterance of the native speech database is recognized by a phone recognizer. Secondly, recognition results are aligned

with their corresponding transcriptions by the dynamic programming. Thirdly, the confusion value of phone *p* with respect to phone *q* is calculated as follows:

$$Conf(p,q) = \frac{M_{pq}}{N_p} \tag{4}$$

where N_p is the total number of phone *p* in transcriptions, and M_{pq} is the times that phone *p* is aligned with phone *q*. Thus, given *N* phones, by iterating phone *p* and *q* though the entire phone set, an $N \times N$ matrix is created. Finally, for each phone *p*, the confusion values are sorted in decreasing order to form the confusion matrix.

It's worth noting that in formula (4), $Conf(p,q)$ becomes the recognition accuracy of phone *p* if *p* equals to *q*. So for each row labeled phone *p*, the first column is usually correspondent with *p* itself, and the most similar phones of *p* are supposed to have low column indices.

3.3.2 Feature Enhancement

With the confusion matrix, extraction process of the PPV presented in Section 3.2 is modified as follows: Firstly, the competitive phone set D_i is now sorted according to the phone orders of p_i in the confusion matrix, $1 \le i \le N$. Then, for each PPV feature, its first *L* elements are differenced to produce derivatives with the *one versus rest* strategy, and the derivatives are appended to the original PPV. Thus, the enhanced posterior probability vector (EPPV) is given by:

$$EPPV = \left[PPV^T \quad d_{12} \quad \cdots \quad d_{mn} \quad \cdots \quad d_{L-1,L}\right]^T \tag{5}$$
$$d_{mn} = v_m - v_n, \quad 1 \le m < n \le L$$

where v_m and v_n are defined in formula (2), and elements d_{mn} are arranged in order of increasing *m* and *n*.

Note that the phone orders in the confusion matrix are used to select subsets of phones for the derivative features. However, the confusion values are not used, for their dynamic ranges vary greatly within phone sets.

3.4. Classifier Building

By using the PPV or the EPPV as the feature, classifiers can be trained to distinguish between correctly and incorrectly uttered phones. Similar to the PDT presented in Section 2.2, classifiers are phone dependent. Likewise, data sparseness problem ought to be considered. In this paper, the following steps are adopted to decrease its impact:

- Robust classifiers, such as the Support Vector Machines (SVM) [7], are chosen.
- Positive samples of one phone are used as negative samples for the others to increase the valid data.
- To reduce the dimension of feature space, for PPV, the competitive phone sets are chosen from the phone classes, and for EPPV, a smaller *L* is preferred.

4. EXPERIMENTS

4.1. Confusable Phones

The aforementioned confusable phones indicate those phones whose corresponding observations tend to have close PP distributions on the same acoustic model. There are two samples illustrated in Figure 2. In part (a), the observations of both phone /s/ and phone /f/ have close PP distributions on model /s/. And in part (b), the observations of phone /in/ and phone /ing/ are close on model /in/. It is observed that they are inseparable with thresholds, and as is indicated by Section 4.3, the PP and PDT methods are barely discriminative on these phones.

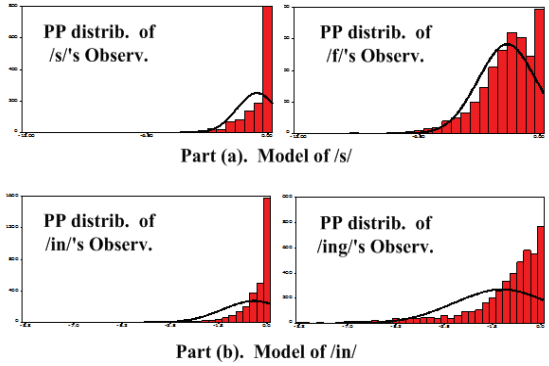

Figure 2. Samples of PP distribution

In this paper, 80 pairs of common confusable phones, which are observed to be intractable in conventional detection tasks, are collected to create the database. Some pairs are listed in Table 1. It's interesting that some of them are also error-prone for human listeners. For example, even for native Chinese, /ing/ and /in/ are hard to differentiate with.

Table 1. Samples of the confusion pairs

Pairs of Initials		Pairs of Finals	
Original	Confused	Original	Confused
/t/	/p/	/ing/	/in/
/s/	/f/	/eng/	/ong/
/n/	/l/	/en/	/an/

4.2. Experiment Setup

There are total 209 native speakers (99 males and 109 females) in our database. 169 speakers (79 males and 90 females) are recorded with original transcriptions to get Dataset-A. The other 40 speakers are recorded with modified transcriptions to get Dataset-B. The modification is carried out to substitute one original phone with the others in its confusable pairs. Finally, there are 55832 valid utterances in Dataset-A, and 12589 ones in Dataset-B. Most of them consist of two syllables in Mandarin.

While Dataset-A is used to tune the model parameters, Dataset-B is used as the test set. Since the modifications of

transcriptions are carried out on confusable pairs, Dataset-B is characterized as the test set for confusable phones.

It's important to know that the references for the testing on Dataset-B are auto-generated during the recording process, for the modifications of transcriptions are traceable.

In our experiment, acoustic models are monophone-model trained from the widely used continuous Mandarin speech database of the national "863" project. SVM models are trained on both the PPV and the EPPV features of Dataset-A with LIBSVM [8]. The L value of the EPPV is set to 5. The confusion matrix is extracted from Dataset-A.

4.3. Performance Comparison

Performance of mispronunciation detection is measured by false alarm rate (FA) and false rejection rate (FR). FA indicates the percentage of correct pronunciations detected as incorrect ones; likewise, FR indicates the percentage of mispronunciations detected as correct ones. For FA and FR varies with different thresholds, Detection Error Tradeoff (DET) curves are plotted. Since FA related errors are more serious for the user satisfaction, practically, systems with a low FA, typically less than 0.1, are favored in use.

Table 2. Detection performance of four methods

FA	FR			
	PP	PDT	PPV+SVM	EPPV+SVM
0.02	0.779	0.669	0.502	0.472
0.04	0.623	0.508	0.360	0.330
0.06	0.508	0.408	0.283	0.262
0.08	0.418	0.344	0.237	0.213
0.10	0.351	0.294	0.195	0.174

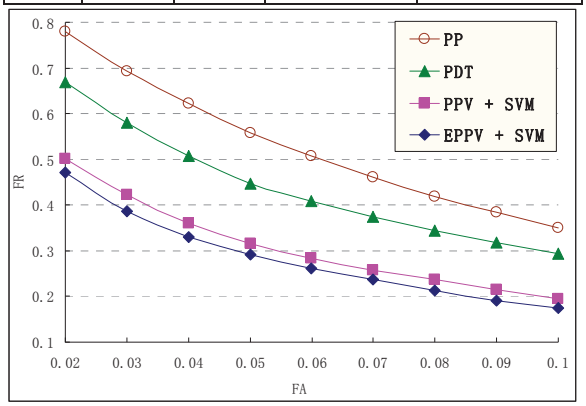

Figure 3. DET curves of four methods

The four aforementioned methods are examined in Table 2 and Figure 3. As is indicated in Table 2, when FA is in the range of 0.02 to 0.1, the PDT reduces FR by 5.7% to 11% (absolute), compared with the unique threshold (PP); and classifiers with the PPV outperform the PDT by 9.9% to 16.7% (absolute); moreover, classifiers with the EPPV outperform the PPV by 2.1% to 3% (absolute), and outperform the PDT by 12% to 19.7% (absolute).

Detailed DET curves plotted in Figure 3 also illustrate the performance gain of our proposed method. Clearly, when FA is in the range of 0.02 to 0.1, classifiers with the PPV reduce FR for more than 10% (absolute), compared with the PDT. Nevertheless, classifiers with the EPPV reduce FR for another 2% (absolute), compared with the PPV. Thus, it is observed that the proposed method is effective for FR reduction in the concerned FA range.

5. CONCLUSION

In this paper, phone-level classifiers are proposed as the detection scheme for the mispronunciation detection on the confusable phones. The confusion matrix is introduced to enhance the performance. The effectiveness of the proposed method are confirmed by the experiments on the utterances of two Mandarin syllables, reducing FR by 12% to 19.7% (absolute) when FA is in the range of 0.02 to 0.1.

In the future, construction of robust classifiers on data with noise, and solution of data sparseness, should also be considered. On the other hand, speaker adaptive training should be incorporated into our framework to provide better statistical models for the detection task.

6. REFERENCES

[1] Witt, S., "Use of Speech Recognition in Computer-Assisted Language Learning", PhD thesis, Cambridge University Engineering Department, Cambridge, UK, 1999.

[2] Franco, H., Neumeyer, L., Kim, Y., Ronen, O., Bratt, H., "Automatic Detection of phone-level mispronunciation for language learning", *in Proc. Eurospeech*, Vol. 2, pp. 851-854, 1999.

[3] Ito, A., Lim, Y., Suzuki, M., Makino, S., "Pronunciation Error Detection Method based on Error Rule Clustering using a Decision Tree", *in Proc. EuroSpeech*, pp. 173-176, 2005.

[4] Zheng, J., Huang, C., Chu, M., Soong, F. K., Ye, W., "Generalized Segment Posterior Probability for Automatic Mandarin Pronunciation Evaluation", *in Proc. ICASSP*, pp.201-204, Hawaii, USA, 2007.

[5] Zhang, F., Huang, C., Soong, F. K., Chu, M., Wang, R., "Automatic mispronunciation detection for Mandarin", *in Proc. ICASSP*, pp. 5077-5080, Las Vegas, Nevada, U.S.A, 2008.

[6] Bouselmi G., Fohr D., Illina I., Haton J.P., "Fully Automated Non-Native Speech Recognition Using Confusion-Based Acoustic Model Integration and Graphemic Constraints", *in Proc. ICASSP*, pp. 345-348, Toulouse, France, 2006.

[7] Vapnik, V.N., "The nature of statistical learning theory", Springer, Second edition, 1995.

[8] Chang, C.-C., Lin, C.-J., "LIBSVM: A Library for Support Vector Machines", At: http://www.csie.ntu.edu.tw/~cjlin/libsvm.