AUTOMATIC PRONUNCIATION ERROR DETECTION BASED ON LINGUISTIC KNOWLEDGE AND PRONUNCIATION SPACE

Shuang Xu, Jie Jiang, Zhenbiao Chen, Bo Xu

Institute of Automation, Chinese Academy of Sciences, Beijing, China, 100080

ABSTRACT

This paper presents a new approach that uses linguistic knowledge and pronunciation space for automatic detection of typical phone-level errors made by non-native speakers of mandarin. Firstly, linguistic knowledge of common learner mistakes is embedded in the calculation of log-posterior probability and the revised log-posterior probability (RLPP) is regarded as the measure of mispronunciation; secondly, a restricted pronunciation space is constructed by using RLPP vectors to describe the characteristics of pronunciation and Support Vector Machine (SVM) classifier is applied into the detection of typical pronunciation errors. Experiments based on a non-native speaker database of mandarin confirm the promising effectiveness of our methods.

Index Terms— Pronunciation error detection, log posterior probability, linguistic knowledge, pronunciation space

1. INTRODUCTION

Recently, Computer Aided Language Learning (CALL) has received a considerable attention in the field of language teaching. Many speech processing and recognition methods have been successfully used to improve the performances of such systems [1-3]. Most of these studies focused on the pronunciation scoring, which allows us to obtain pronunciation quality scores for sentences or phrases, with grading consistency similar to that of human experts. While pronunciation scoring plays an essential role in CALL system, providing more detailed feedbacks on specific pronunciation errors is also very important to help correct or improve pronunciation. This paper focuses to improve the performance of automatic detection of pronunciation errors, especially the typical phone-level errors made by non-native speakers of mandarin.

Phone-level pronunciation errors are the errors concerned with the distinguishable sound units of speech, i.e. initials or finals for mandarin. Most of the typical phone-level errors are systematic errors, which are governed by the linguistic structure of learner's native language and target learning language. In mandarin teaching, these systematic errors have been regarded as learning points and summarized by human tutors. The prior linguistic knowledge of pronunciation errors can be used in CALL system to improve performance [4].

In the detection of mispronunciation, the most commonly used method is log-posterior probability (LPP) based method, which uses LPP as the measure of mispronunciation, with uniform or phone-dependent threshold strategy [5] [6]. However, this method was shown to have poor performance in the detection of systematic errors—a difficult task that involves discrimination between close pronunciations. In order to improve the detection accuracy, a revised log-posterior probability (RLPP) based on prior linguistic knowledge of pronunciation errors was proposed and applied into the detection of systematic errors.

Motivated by the methods used in [7], we use RLPP vectors to construct pronunciation spaces restricted by prior linguistic knowledge and use Support Vector Machine (SVM) classifiers to identify typical pronunciation errors. In a restricted pronunciation space, RLPP vectors are used as features to represent learner's variable pronunciation and SVM classifier are used to catch the discriminative information embedded in the RLPP vectors.

The paper is organized as follows. Section 2 describes linguistic knowledge used in our system. Section 3 presents the calculation and application of RLPP. Section 4 describes the implementation of restricted pronunciation space. The experiment results and analysis are given in Section 5. Section 6 summarizes the paper.

2. LINGUISTIC KNOWLEDGE OF ERRORS

In the past, the development of speech processing and recognition technologies for CALL systems was solely left in the hands of speech engineers and did not make use of the experiences of human tutors, who draw from their knowledge of typical learner mistakes to understand learner speech and offer effective instructions. Here, we attempt to apply the prior linguistic knowledge into our CALL system to improve the detection accuracy of typical pronunciation errors.

According to the second language learning theory [8], the typical pronunciation errors made by non-native speakers are mostly produced by the influence of their mother language and can be predicted statistically.

Therefore, the linguistic knowledge used in our CALL system not only includes the knowledge of human tutors, but also includes error patterns derived from large amounts of pronunciation data. The used linguistic knowledge is listed in Table.1.

Table.1. Linguistic knowledge of pronunciation errors

#	Description
1	Confusion of fricatives and affricatives
2	Confusion of labiodentals and bilabials
3	Confusion of flat tongue and raised tongue
4	Confusion of alveolo-palatals and retroflexs
5	Replacement of unvoiced consonants with voiced consonants
6	Replacement of aspirated consonants with non-aspirated consonants
7	Replacement of [ʐ] with [r]
8	Replacement of [w] with [v]
9	Replacement of [x] with [ɤ]
10	Confusion of [əu] and [uə]
11	Confusion of [ɤ] and [i]
12	Replacement of [u] with [o] or [əu]
13	Confusion of alveolar nasals and velar nasals
14	Adding medium to a vowel
15	Omission of medium of a vowel
16	Omission of coda of a vowel

In Table.1, the top 9 error patterns (1~9) are concerned with initial errors and the left (10~16) are final errors. For each initial or final, there are likely to have more than one error patterns to be associated with.

3. RLPP BASED ERROR DETECTION

This section describes the calculation of RLPP and its application in pronunciation error detection.

3.1. Calculation of RLPP

In many CALL systems, LPP has been regarded as the measure of pronunciation quality because of its robustness to variations in speaker identity and acoustic channel. Under statistical speech recognition framework, LPP of phone segment q_i is calculated as:

$$LPP_i = \frac{1}{T_i}\log\left(\frac{\Pr ob(o_i \mid q_i)}{Max_{j\in Q}(\Pr ob(o_i \mid q_j))}\right) \quad (1)$$

Where Q is the model set, T_i is the total frame of phone q_i, $\Pr ob(o_i \mid q_i)$ is the likelihood of HMM model q_i given the observation vector o_i.

In Formula (1), the denominator of LPP is obtained from the entire phone loop network, which introduces large confusions among different HMM models and results in a significant reduction in performance. In a speech recognizer, the confusions among HMM models can be eliminated to a large extent by the introduction of language model. Motivated by the idea of recognizer, we use the linguistic knowledge described in section 2 to revise the calculation of LPP from Formula (1) to Formula (2). The revised LPP is called RLPP:

$$RLPP_i = \frac{1}{T_i}\log\left(\frac{\Pr ob(o_i \mid q_i)}{Max_{j\in Q_i}(\Pr ob(o_i \mid q_j))}\right) \quad (2)$$

The difference between LPP and RLPP is the calculation of the denomination. The model set Q used by LPP includes all the phone models while Q_i used by RLPP only includes the correct pronunciation model and the typical mispronounced models of phone q_i. The choice of elements in Q_i lies on the corresponding linguistic knowledge of phone q_i. By utilizing a restricted network in denomination, RLPP can diminish the influence of model confusions and capture typical pronunciation errors with better performance.

3.2. RLPP based error detection

With RLPP, a system for detection of typical pronunciation errors is easily implemented. A block diagram of such system can be seen in Fig.1.

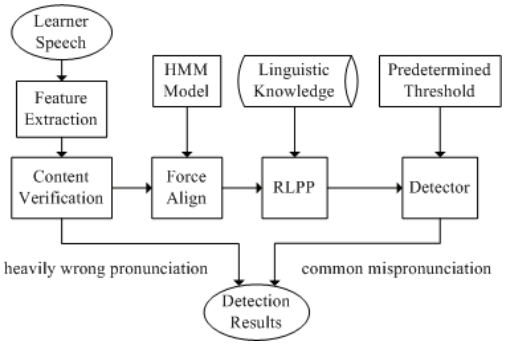

Fig.1. Block-diagram of RLPP based method

In Fig.1, the feature extraction module converts the learner speech to a sequence of frames with perceptual linear predictive (PLP) coefficients. The content verification module uses speech recognition to verify the content of learner speech. If the content is not consistent with the text which the learner is asked to read out, the speech is rejected and regarded as heavily wrong pronunciation. On the contrary, if the speech is not rejected, it will be processed in the following module. The force align module is used to convert input speech to different phone segments based on Viterbi decoding algorithm. In RLPP module, the individual RLPP score is calculated for each phone as defined in Formula (2). Finally, in detector module, phone-dependent threshold is applied to each RLPP score and the phone segment with RLPP score lower than corresponding threshold is regarded as a mispronounced segment.

4. ERROR DETECTION BASED ON PRONUNCIATION SPACE

In speech recognition, we use basic-phone models to describe different pronunciations. This kind of model space is not suitable to our task because of the diversity of mispronunciation, which may be a complete replacement of a phone with another phone or only a part replacement. In this section, by using RLPP vectors, we construct a pronunciation space for each phone to describe its variable pronunciation. The space is called restricted pronunciation space (RPS), so named because it is restricted by linguistic knowledge, not only in the calculation of RLPP, but also in the elements of space.

In addition, since the goal of pronunciation error detection is to classify correct pronunciation and incorrect pronunciation, it can be regarded as a direct binary classification problem. Motivated by the recent development in speaker verification, SVM is used as classifier to detect pronunciation errors, which seems well suited to binary classification.

Next we explain our method that uses RLPP vectors and SVM classifier in RPS to find typical pronunciation errors.

Firstly, RLPP vectors are extracted. For phone segment q_i , the RLPP vector is represented as: $\{RLPP_i^0, RLPP_i^1, RLPP_i^2, \cdots RLPP_i^{N_i}\}$, let $RLPP_i^0$ be the RLPP given the correct pronunciation model of phone q_i , $RLPP_i^1$, $RLPP_i^2$,…, $RLPP_i^{N_i}$ be the RLPPs given the different mispronounced models of phone q_i , N^i be the number of error types concerned with phone q_i . The correct pronunciation model and different mispronounced models are pre-trained with correct pronunciation data and variable mispronunciation data respectively. These models construct the model set Q_i in Formula (2) and the RPS of phone q_i .

Secondly, with RLPP vectors, SVM classifier for every phone is trained to catch the discriminative information embedded in the RLPP vectors. The process of training is shown in Fig.2. Training of SVM classifier requires information about positive data and negative data. Therefore, the training data should be split in a manner such that every phone is associated with a pair of data sets, one for positive data and the other representing the negative data. The positive data is obtained from RLPP vectors of speech segments marked with 1 by human tutors (correct pronunciation) while the negative data consists of RLPP vectors of speech segments marked with 0 by human tutors (mispronunciation).

Finally, using RLPP vectors as features, pre-trained SVM classifier is used to classify correct pronunciation and incorrect pronunciation of a phone.

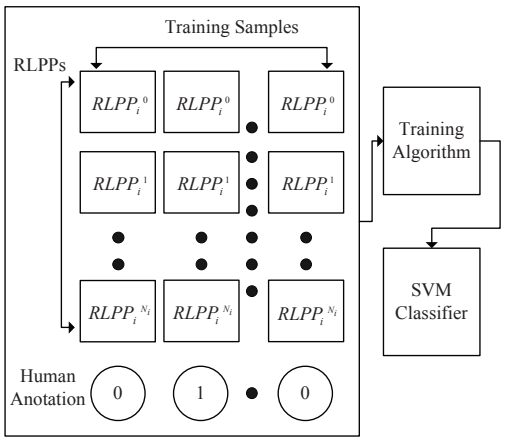

Fig.2. Training of SVM classifier model with RLPP vectors

5. EXPERIMENTS AND RESULTS

5.1. Database

The following experiments are performed on a test database which is carefully designed in order to be consistent with Putonghua-Shuiping-Ceshi (PSC) — a national test to evaluate the proficiency of spoken mandarin. The database consists of 1585 words (6140 phones) pronounced by non-speakers of mandarin (50 female and 50 male). About 2950 phones are pronounced incorrectly and most pronunciation errors are typical errors as listed in Table.1.

The speech used for training restricted pronunciation space models and SVM classifier models are from a standard mandarin database spoken by native speaker and a mandarin database spoken by non-native speaker, in which a lot of pronunciation errors have been annotated by human tutors.

5.2. Experiments

5.2.1. Performance measures

In our pronunciation error detection task, two measures are considered to compare the performance of different methods: false alarm rate (FA) and false rejection rate (FR). They are defined as follows:

$$FA = \frac{Num_{rw}}{Num_{right}}, \quad FR = \frac{Num_{wr}}{Num_{wrong}} \tag{3}$$

Where Num_{rw} is the number of phones whose pronunciation are correct but identified as mispronunciation, Num_{wr} is the number of phones whose pronunciation are incorrect but identified as correct pronunciation, Num_{right} is the total number of phones with correct pronunciations and Num_{wrong} is the total number of phones whose pronunciations are incorrect.

5.2.2. Effect of introduction of linguistic knowledge

To validate the effectiveness of linguistic knowledge, we compare the performance of two error detection methods: LPP based method and RLPP based method. LPP based method is similar to RLPP based method, which is demonstrated in Fig.1. The main difference between the two methods is in the utilization of linguistic knowledge. LPP based method doesn't use any linguistic knowledge while RLPP based method use linguistic knowledge as listed in Table.1.

The result of the experiment is shown in Fig.3. It is very clear that RLPP based method makes better performance than LPP based method. For all FA regions, it can reduce FR greatly, e.g. for 0.05 FA, FR reduces from 0.494 to 0.228. It is what we expect in a real CALL system.

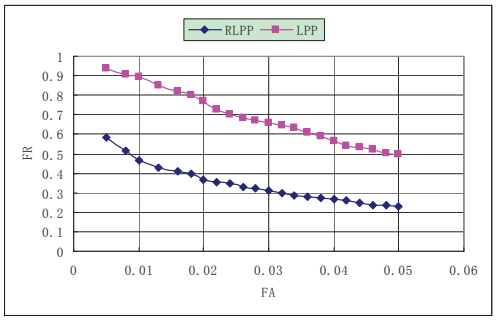

Fig.3. Performance curves with LPP and RLPP

5.2.3. Effect of introduction of RPS

With introduction of RPS, the phones whose pronunciation is between correct pronunciation and incorrect pronunciation can be represented more exactly. The performance of RPS based method for error detection is shown in Fig.4. As we observe, this method can further improve the accuracy of error detection, e.g. for 0.05 FA, FR reduces from 0.228 to 0.175.

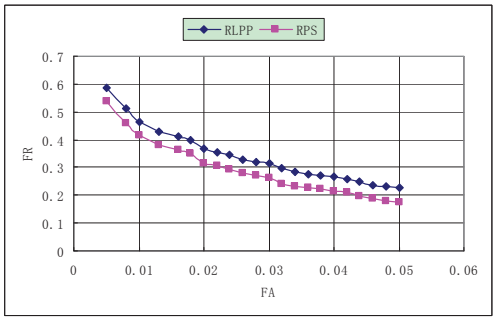

Fig.4. Performance curves with RLPP and RPS

5.2.4. Detailed experimental results

The detailed results of above experiments are given in Table.2. From Table.2, performances of FR on the same condition of FA can be clearly compared among the three methods mentioned above. As we observe, RPS method can achieve the best performance for all FA regions.

Table.2. Detailed performances of experiments

FA	FR		
	LPP	RLPP	RPS
0.05	0.494	0.228	0.175
0.04	0.566	0.266	0.216
0.03	0.657	0.313	0.264

6. CONCLUSION

In this paper, we focus on the detection of pronunciation errors that are most commonly made by non-speaker of mandarin. Approaches from two aspects are proposed to improve the performance of our system: revising the calculation of LPP using linguistic knowledge of common learner mistakes; constructing a restricted pronunciation space based on RLPP vectors and applying Support Vector Machine (SVM) classifiers into the mispronunciation detection. Experiments based on a non-native speaker database of mandarin confirm the promising effectiveness of our methods.

7. REFERENCES

[1] C. Cucchiarini, H. Strik, and L. Boves, "Different Aspects of Expert Pronunciation Quality Ratings and Their Relation to Scores Produced by Speech Recognition Algorithms", *Speech Communication*, Elsevier, pp. 109-119, 2000.

[2] L. Neumeyer, H. Franco, M. Weintraub, etc, "Automatic Text-Independent Pronunciation Scoring of Foreign Language Student Speech", in *Proc. ICSLP*, Philadelphia, pp. 1457-1460, 1996.

[3] Kei Ohta and Seiichi Nakagawa, "A Statistical Method of Evaluation Pronunciation Proficiency for Japanese Words", in *Proc. Interspeech*, Lisbon, pp. 2233-2236, 2005.

[4] S. M. Witt and S.J. Young, "Language Learning Based on Non-Native Speech Recognition". In *Proc. Eurospeech*, Greece, pp. 633-636, 1997.

[5] H. Franco, L. Kim, etc, "Automatic Detection of phone-level mispronunciation for language learning", in Proc. Eurospeech, pp. 851-854, 1999.

[6] P. Fuping, Z. Qingwei, Y. Yonghong, "Mandarin Vowel Pronunciation Quality Evaluation by A Novel Formant Classification Method and Its Combination with Traditional Algorithms", in *Proc. ICASSP*, pp. 5061-5064, 2008.

[7] V. Wan, "Speaker Verification using Support Vector Machines", Ph. D Thesis, University of Sheffield, United Kingdom, 2003.

[8] J. Jenkins, "The learning theory approach", *Psycholinguistics: A Survey of Theory and Research*, pp. 20-35, 1954.