Automatic pronunciation error detection of Chinese based on SVM and structural features *

Tongmu Zhao (University of Tokyo), Akemi Hoshino (Toyama National College of Technology), Masayuki Suzuki, Nobuaki Minematsu, Keikichi Hirose (University of Tokyo)

1 Introduction

Pronunciation errors are often made by learners of a foreign language. Especially when the target language contains some phonemes that are not found in learners' native language, learns tend to replace these phonemes with ones existing in their native language. Automatic detection of these errors is an essential technique in CALL systems [1], which can accelerate foreign language learning.

A novel and structural model of pronunciation has been recently proposed, which works efficiently to discard the non-linguistic aspects of speech, which are irrelevant to pronunciation assessment, and keep the linguistic aspects well [2]. Besides, this structure model has been applied to speech recognition [3], speech synthesis [4], overall pronunciation scoring [5], and dialect-based speaker clustering [6]. This paper reports our first trial to apply our structural model to phoneme error detection.

2 Materials

We create two native databases by asking native speakers to read given sentences. While one is error free, named Chinese Read by Natives (CRN), the other contains phoneme errors intentionally introduced by the speakers, named Chinese Read by Natives with Errors (CRN-E).

Both the databases are constructed in the following way. At first, 17 sentences are selected from a Chinese textbook [7] as reading material. In CRN database, 4 Chinese speakers are asked to read the material to form CRN database.

Then, through discussion with Chinese teachers, 8 target phonemes are selected, which are the most problematic and difficult phonemes for Japanese learners to pronounce correctly. Further, for each target phoneme, its competitive one is selected, which is often substituted incorrectly by Japanese students for the target phoneme. Table 1 shows the 8 phoneme pairs. For example, when Japanese wants to pronounce /sh/, he may pronounce /x/ instead.

Table 1 Eight target phonemes and their competitive ones

Targets	zh	ch	sh	v	er	ing	eng	ang
Competitive	j	q	x	u	a	in	en	an

Because of difficulty of preparing a labeled non-native speech database, [8] prepared data of phoneme errors by changing phoneme-based transcripts. [8] shows technical effectiveness and validity of this "artificial" preparation. In our study, artificial data is created by asking natives to read sentences with intentional errors. 48% of the instances of the 8 target phonemes are replaced by their competitive ones in the reading material. 9 native speakers are asked to read this material. Each speaker read 3 times per sentence. Their utterances formed a database of CRN-E.

Moreover, 30 speakers in NICT database [9] are used for training native phoneme HMMs and determining GOP thresholds. Summary of the 3 databases is shown in Table 2. "#M/F" represents the number of male and female speakers. "#U" represents the number of utterances.

Table 2 Summary of the 3 databases

Name	#M/F	#U	Usage
CRN	2/2	80	Teachers' structural model
CRN-E	5/4	459	Testing data for the three error detection methods
NICT	10/10	5000	Native HMM training
	5/5	2500	GOP thresholds

*SVM と構造表象に基づく中国語発音誤りの自動検出, 趙 童牧(東京大学), 星野 朱美(富山高等専門学校), 鈴木 雅之, 峯松 信明, 広瀬 啓吉(東京大学)

3 Methods

3.1 Goodness of Pronunciation (GOP) and Likelihood Ratio (LR)

GOP calculates the likelihood ratio that a phoneme realization corresponds to the phoneme that should have been spoken [10]. A GOP score of phoneme /x/ is a posterior probability of the phoneme given corresponding speech segment, which is approximated by (1). O is the speech segment obtained for /x/.

$$\mathrm{GOP}(x,O) = P(x|O) \approx \log\left(\frac{P(O|x)}{\max_{y\in Q} P(O|y)}\right) \qquad (1)$$

By using correctly pronounced data and incorrect data, distribution of GOP scores of correct pronunciation and that of errors can be obtained. By using target phoneme dependent threshold α, we can do error detection [8]. If GOP(x,O)≥α, segment O is judged as correct and otherwise not.

In data collection, the correspondence between intended phonemes and substituted phonemes is assumed to be fixed as in [8]. Hereafter, we use /x/ as intended phoneme and /y/ as substituted phoneme. In preparing the CRN-E database, we used the information of /y/. But GOP does not exploit this information, which makes GOP not a fair comparison with our SVM method. So, LR is also tested [11]. An LR score of phoneme /x/ is calculated by taking the absolute difference of the log probability calculated through the forced alignment as /x/ and that as /y/. In error detection, if the LR score is higher than 0, O is judged as correct and otherwise, not.

$$\mathrm{LR}(x,y,O) = \log\left(\frac{P(O|x)}{P(O|y)}\right) \qquad (2)$$

3.2 Structural features

This section explains the process of extracting structural features. An utterance is represented by a sequence of feature vectors, which is then converted into a distribution sequence. Distance between every distribution pair is calculated as root of Bhattacharyya distance. A full set of distances, i.e. distance matrix, is used to represent this utterance [2]. Suppose that a teacher and a student read the same sentence and both the utterances are converted into distance matrices of $\{S_{ij}\}$ and $\{T_{ij}\}$. The structural deviation related to phoneme i is calculated by (3), which quantifies the magnitude of structural difference as for phoneme i between the teacher and the student [5]. M is the number of distributions, which can be phonemes or states.

$$D(S,T,i) = \sum_{j=1}^{M} \left\| \frac{S_{ij} - T_{ij}}{S_{ij} + T_{ij}} \right\| \qquad (3)$$

3.3 Support Vector Machine

One problem in structural features is that not all the elements in {Dij} are supposed to contribute to (3) with the same importance. When multiple phonemes are incorrectly pronounced in a sentence, the distance to one of the erroneous phonemes will impede the detection performance. So, we introduce SVM to solve this problem. Let x_i represent a structural difference vector of phoneme i, and y_i represent a 1/0 label of x_i. Given a training set of instance-label pairs of (x_i,y_i), the SVM is obtained by solving the following problem [12]: x_i is mapped into a hyperplane by function ∅, b is the bias term of the hyperplane. C(>0) is the penalty parameter of the error term ε_i. W is the weight vector of x_i.

$$\begin{aligned} \text{subject to} \quad & y_i(W^T \emptyset(x_i) + b) \geq 1 - \varepsilon_i \\ \min_{w,b,\varepsilon} \quad & \frac{1}{2} W^T W + C \sum_{i=1}^{M} \varepsilon_i \\ & \varepsilon_i \geq 0 \end{aligned}$$

3.4 Performance measures

Error detection produces four basic outcomes: #correct acceptances (CA), #correct rejections (CR), #false acceptances (FA), #false rejections (FR) [8]. The performance of an error detection algorithm can be measured as scoring accuracy, S A = (CA+CR)/(CA+CR+FA + FR) [8]. Besides, we define Precision of CA (PCA), Precision of CR (PCR) [8], False Rejection Rate (FRR), False Acceptance Rate (FAR), Average Error Rate (AER) [13] as follows: PCA = CA /(CA + FA), PCR = CR / (CR + FR), FAR = FA / (CR + FA), FRR = FR / (CA + FR), and AER = (FAR + FRR)/2.

4 Experiments and Results

4.1 GOP and LR-based error detection

In the NICT database, artificial pronunciation errors are created by changing the transcript as in [8]. Some instances of the phonemes in the

second row of Table 1 are replaced by their competitive phonemes in the first row. Then, GOP scores of correct pronunciations and those of mispronunciations are calculated separately. In Fig.2, the GOP distribution of /sh/ (correct pronunciation) is drawn in blue, while the GOP distribution of /sh/ with error (real pronunciation is /x/) is drawn in red. We set the threshold so as to minimize the classification error. The thresholds of all the target phonemes are obtained from their corresponding distributions.

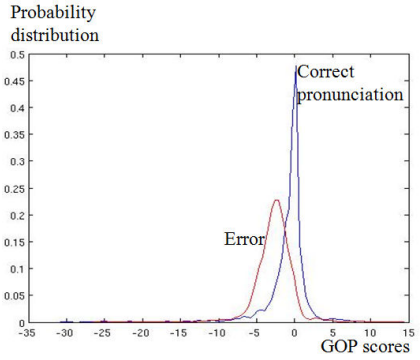

Fig. 2 Probability distribution of /sh/ GOP scores

Finally, each instance of the target phonemes of the testing utterances is judged as correct or not using GOP and LR. Table 3 shows our results and the results of other studies just as reference. From the table, SA is 0.82 in [8], where the target language is Dutch and it is 0.60 in [13] where it is Mandarin. This large difference is due to language difference. Comparing the performance in the same language, our SA is similar to [13], while we have a better result of FRR but a much worse result of FAR. Here, due to differences of other experimental conditions, we do not discuss the performance difference to [13] any further.

Results of LR-based error detection in CRN-E databases are also shown in Table 3. SA of LR is much higher than the result of GOP. Specifically, FAR is reduced greatly in LR-based error detection but FRR increased slightly.

Table 3 GOP and LR-based error detection

	Our study		[13]	[8]
Language	Mandarin			Dutch
Methods	GOP	LR	GOP	GOP
SA	0.59	0.75	0.60	0.82
PCA	0.58	0.77		0.82
PCR	0.65	0.73		0.81
FAR	0.75	0.26	0.42	
FRR	0.12	0.24	0.24	
AER	0.43	0.25	0.33	

4.2 SVM with structural features

When using SVM with structural features in error detection, firstly, structural features should be extracted. Forced alignment is firstly done using the NICT phoneme HMMs and, then, using the boundary information, Viterbi training is done to train an HMM only for that utterance. Each utterance of each student and that of each teacher is converted to its HMM and its distance matrix. Here, each phoneme instance is treated separately. An M×M distance matrix has to be estimated for an utterance, where M is the number of phoneme instances in the utterance.

As for SVM, LIBSVM [14] is adopted. The CRN-E database is divided into training part and testing part. For each sentence, the teachers' matrix is obtained as the average matrix among the four teachers. Then, equation (3) is used to obtain the structural deviation of each phoneme instance in each of the students' utterances.

Then, a leave-one-out cross-validation test is done. For a sentence, we set 1 speaker out of 9 as testing data and the other speakers as training data for SVM with the linear kernel. By changing speaker assignment, we used all the speakers as testing speakers. Table 4 is the average performance over the 9 experiments. We can see that the proposed SVM with structural features work better than the baseline LR-based method in all measures. Especially, PCR is increased by 29.4% and FRR is decreased by 81.5%.

Table 4 Comparison of error detection using LR-based and SVM with structural features

	LR	SVM+str. features	Comparison
SA	0.75	0.88	+17.3%
PCA	0.77	0.85	+9.2%
PCR	0.73	0.94	+29.4%
FAR	0.26	0.21	-20.8%
FRR	0.24	0.04	-81.5%
AER	0.25	0.13	-49.8%

4.3 Use of partial structural matrices

In the CRN-E database, error position is fixed. So, in a sentence, some instances of a target phoneme are always correct and the others of the same target phoneme are always incorrect. This strategy of fixed assignment will be beneficial for SVM training. So, we test SVM using no edge

related to the 8 target phonemes, where a target phoneme instance is evaluated only with its relations to all the non-target phonemes. Results are shown in Table 5. Further, we ran another test to evaluate the robustness of the structure-based SVM experimentally. Here, cross-gender error detection is done, whose results are also shown in Table 5. They are results of the two cases where training speakers for SVM are only 3 males and 3 females, respectively. Each performance measure shows very similar scores. This indicates very high robustness of our proposed method.

Table 5. Results of using partial matrices

Test types	Cross-gender		Leave-one-out
	3 male	3 female	
SA	0.84	0.84	0.83
PCA	0.80	0.80	0.78
PCR	0.91	0.92	0.92
FAR	0.27	0.28	0.31
FRR	0.06	0.05	0.05
AER	0.16	0.17	0.18

5 Conclusion

In this paper, the most problematic phonemes for Japanese learners of Chinese are defined and error detection for these phonemes is investigated. We designed two new databases, one of which included intentional phoneme errors generated by natives. Three methods of error detection are tested, where partial matrices as well as complete matrices are examined. Our proposed SVM with structural features works much better than both of the GOP-based and LR-based baseline methods.

One problem is that the errors are limited to the 8 target phonemes. SVM's good performance may be due to the assumption that the non-target phonemes are always pronounced correctly. One possible solution is collecting a large amount of learners' data with labels, with which unreliable edges will be ignored through SVM training automatically. We are also planning to use real Japanese learners' data in the near future.

References

[1] M. Eskenazi, "An overview of spoken language technology for education," Speech Communication, 51, 832-844, 2009

[2] N. Minematsu et al., "Speech structure and its application to robust speech processing," Journal of New Generation Computing, 28, 3, 299-319, 2010

[3] M. Suzuki et al., "Discriminative reranking for LVCSR lever-aging invariant structure," Proc. INTERSPEECH, 2012 (to appear)

[4] S. Saito et al., "Structure to speech conversion speech generation based on infant-like vocal imitation," Proc. INTERSPEECH, 1837–1840, 2008

[5] M. Suzuki et al., "Integration of multilayer regression with structure-based pronunciation assessment," Proc. INTERSPEECH, 586-589, 2010

[6] X. Ma et al., "Structural analysis of dialects, sub-dialects, and sub-sub-dialects of Chinese," Proc. INTERSPEECH, 2219-2222, 2009

[7] Chinese reading materials: Xingren, University of Tokyo Faculty of Arts Committee, 2008

[8] S. Kanters et al., "The goodness of pronunciation algorithm: a detailed performance study," Proc. SLaTE, CD-ROM, 2009

[9] NICT Chinese database: http://alagin.jp/

[10] S. M. Witt et al., "Phone-level pronunciation scoring and assessment for interactive language learning," Speech Communication, 30, 95-108, 2000

[11] H. Franco, et al., "Combination of machine scores for automatic grading of pronunciation quality," Speech Communication, 30, 121-130, 2000

[12] C. W. Hsu et al., A practical guide to support vector classification. Tech. rep., Department of Computer Science, National Taiwan University, 2003

[13] Y.B. Wang, "Improved Approaches of Modeling and Detecting Error Patterns with Empirical Analysis for Computer-Aided Pronunciation Training," Proc. ICASSP, 5049-5052, 2012

[14] C. Chang et al., LIBSVM: a library for support vector machines, 2001.