Automated Pronunciation Scoring for L2 English Learners

1. Introduction

This study aims at developing an automated pronunciation scoring method for second language learners of English (Hereafter, L2 learners) using both confidence scoring and classifiers. The pronunciation errors have been detected using the confidence measure from speech recognition [Franco et al., 1997, Neumeyer et al., 2000, Witt and Young, 1997, Witt, 1999]. However, the accuracy of the assessment based on the confidence scores is not always high. In contrast to the confidence scoring method, which has been the traditional method for pronunciation scoring, the classifier method can be trained using a small amount of data, and achieve high accuracy in the binary classification of phonetic features. Due to these advantages, [Truong et al., 2004] developed a classifier model which distinguishes the correct sounds from incorrect sounds. [Truong et al., 2004] used different acoustic feature vectors according to the target L1 phoneme, but implementing a per phone feature extraction algorithm is a difficult and time-consuming task. [Mark Hasegawa-Johnson and Wang, 2005] used the same acoustic features with speech recognition for all phonemes, but they could achieve high accuracy of a binary feature classification by extracting the feature vector from the appropriate time interval according to the phonemes. In this study, classifiers were trained for the specific English phonemes where L2 learners make frequent errors. Similar to [Mark Hasegawa-Johnson and Wang, 2005], the same acoustic features with speech recognition was used. The pronunciation scoring method based on the combination of two scores improve the accuracy of pronunciation error detection.

2. Training of Confidence Score

The confidence score measures how closely the utterance of the L2 speaker matches an L1 target phoneme. Mismatches result in low confidence scores which provide a profile of the speakers' production errors. The confidence score is calculated based on the following ways: (1) An English acoustic model was trained using 1997 HUB4 data [Pallet, 1997] (broadcast news data). (2) The start time and end time of target phonemes were estimated based on the time aligned transcription. (3) Phone lattice was generated using a triphone acoustic model. (4) For each phone in word, confidence score is calculated using the following formula .

$$C_{MAP}(c_i|x) = P(c_i|x) = \frac{P(x|c_i) \times P(c_i)}{P(x)}$$
$$where\ c_i = \text{speaker}'\text{s target phoneme}$$
$$x = \text{given acoustic signal}$$

The confidence score of each phoneme was used as its pronunciation score. If the pronunciation score was lower than threshold, the phoneme was considered as an error.

3. Training of Classifiers

For the English phonemes where L2 learners make frequent errors, their possible substitution patterns were collected from [Swan and Smith, 2002]. SVMs were trained in order to distinguish each L2 phoneme from it's possible substitution patterns. SVMs were trained using TIMIT data (a wideband read speech corpus). Total 2310 sentences from 450 speakers were selected and used for training of SVM. For example, if the target English phoneme is [f], and its potential substitution pattern is [p], then [f] was classified into positive case, while [p] was classified into negative case, and a SVM classifier was trained in order to distinguish [f] from [p]. For each pair, the same numbers of positive samples and negative samples were used for training. SVMs in this study are based on the acoustic feature vector including PLPs (12 PLP coefficients, energy, their deltas and acceleration, computed once/10ms with a 25ms window), formants(f1, f2 formant frequency). SVMs were trained using RBF kernels using the lightSVM toolkit [Joachims, 1999].

4. Experiments

The method was tested on the artificial L1 data instead of the rated speech corpus since no publicly available L2 corpus with phone accuracy score exist. The method was tested on the Buckeye Corpus of conversational speech [Pitt et al., 2005]. The corpus includes spontaneous speech of 40 native English speakers with the diverse ages. 400 sentences from 20 speakers were randomly selected and used as a training and test data.

10 phonemes where Korean/Chinese/Spanish speakers make frequent pronunciation errors were selected and, their possible substitution patterns were found from [Swan and Smith, 2002]. Table 1 shows partial examples of L2 phonemes and their possible substitutions of L1 phonemes (Hereafter, L1 substitution phonemes). The words including L1 substitution phonemes were changed to contain L2 target phonemes. Thus, in point of the dictionary view, L1 speakers make pronunciation errors in those words. In this way, the test data included the artificial pronunciation errors.

Table 1: L1 and L2 Phoneme Pairs used in Artificial Data

L2 target phoneme	L1 substitution phoneme	original word	changed pronunciation
æ	e	seven	s æ v ɪ n
v	b	baby	v eɪ b ɪ

In order to combine confidence score and SVM score, the data were classified into three parts (training/development/test), and a SVM was trained. The ranges of confidence scores vary according to the phonemes, the development data was used in order to find the appropriate threshold for each phoneme. Table 2 summarizes the size of the training/development/test corpus.

Table 2: Training and Test Data for Pronunciation Scoring

Train			Development			Test		
speakers	sentences	samples	speakers	sentences	samples	speakers	sentences	samples
15	300	9626	15	300	9626	10	200	4902

5. Results and Discussion

The performance of the algorithm was evaluated using f-score measure. Table 3 shows f-scores at phoneme specific thresholds. The point where the method achieved the highest f-score in the development test data was determined as a threshold.

The confidence score shows higher f-score for consonants [dh,v,f] and vowel [ih], while SVM scores shows higher f-score for vowel [ae,ao]. By combining two scores, f-scores of all phonemes except [ih] were improved. In average, there was 3% improvement (relative 17%) in f-score. This results shows that combining SVM and confidence score can achieve further improvement in pronunciation error detection. Furthermore, in contrast to the confidence model, which does not explain what the incorrect phonemes are like, SVM can provide the acoustic characteristics of the incorrect phone. This information can be used as a key to provide the valuable feedback to correct the error. Based on this information, the accurate feedbacks how to correct the pronunciation error can be provided to L2 learners.

Table 3: F-scores for each phoneme

F-score	ae	ao	ih	uh	th	dh	f	v	total
Confidence	0.70	0.80	0.86	0.78	0.88	0.84	0.93	0.86	0.83
SVM	0.77	0.88	0.82	0.78	0.87	0.77	0.90	0.84	0.83
Combined	0.78	0.88	0.85	0.79	0.91	0.85	0.95	0.88	0.86

References

[Franco et al., 1997] Franco, H., Neumeyer, L., Kim, Y., and Ronen, O. (1997). Automatic pronunciation scoring for language instruction. In *ICASSP 97*, pages 1471–1474.

[Joachims, 1999] Joachims, T. (1999). MIT-Press.

[Mark Hasegawa-Johnson and Wang, 2005] Mark Hasegawa-Johnson, James Baker, S. B. K. C. E. C. S. G. A. J. K. K. K. L. S. M. J. M. K. S. and Wang, T. (2005). Landmark-based speech recognition: Report of the 2004 johns hopkins summer workshop. In *Automatic Speech Recognition and Understanding Workshop*. ICASSP.

[Neumeyer et al., 2000] Neumeyer, L., Franco, H., Digalakis, V., and Weintraub, M. (2000). Automatic scoring of pronunciation quality. In *Speech Communication*, pages 88–93.

[Pallet, 1997] Pallet, D. (1997). Overview of the 1997 darpa speech recognition workshop. In *DARPA Speech Recognition Workshop*. DARPA.

[Pitt et al., 2005] Pitt, M., Johnson, K., Hume, E., Kiesling, S., and Raymond, D. (2005). Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, pages 90–95.

[Swan and Smith, 2002] Swan, M. and Smith, B. (2002). *Learner English*.

[Truong et al., 2004] Truong, K., Neri, A., Cuchiarini, C., and Strik, H. (2004). Automatic pronunciation error detection: an acoustic-phonetic approach. In *the InSTIL/ICALL Symposium*, pages 135–138.

[Witt, 1999] Witt, S. (1999). *Use of the speech recognition in computer-assisted language learning*. Unpublished dissertation, Cambridge University Engineering department, Cambridge, U.K.

[Witt and Young, 1997] Witt, S. and Young, S. (1997). Performance measures for phone-level pronunciation teaching in CALL. In *the Workshop on Speech Technology in Language Learning*, pages 99–102.