Landmark-based Automated Pronunciation Error Detection

Su-Youn Yoon, Mark Hasegawa-Johnson, and Richard Sproat

[1]Educational Testing Service, Princeton, NJ 08541, USA
[2]University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA
[3]Oregon Health and Science University Portland, OR 97239, USA

syoon@ets.org, jhasegaw@illinois.edu, rws@xoba.com

Abstract

We present a pronunciation error detection method for second language learners of English (L2 learners). The method is a combination of confidence scoring and landmark-based Support Vector Machines (SVMs). Landmark-based SVMs were implemented to specialize the method for the specific phonemes with which L2 learners make frequent errors.

The method was trained for the difficult phonemes for Korean learners and tested on intermediate Korean learners. In the data where distortion errors (non-phonemic errors) occupied high proportion, SVM method achieved significantly higher F-score (0.67) than confidence scoring (0.60). However, the combination of two methods without the appropriate training data did not lead to improvement.

Even for intermediate learners, a high proportion of errors (40%) was related to these difficult phonemes. Therefore, the method specialized for these phonemes will be beneficial for both beginners and intermediate learners.

Index Terms: automated pronunciation error detection, computer aided pronunciation training system, confidence score, landmark-based SVMs

1. Introduction

This study aims at developing an automated pronunciation error detection method for second language learners. Many second language learners (L2-learners) have difficulty in both perceiving and producing phonemes that do not exist in their native language.

In order to effectively train pronunciation, the L2 learner first needs a diagnostic followed by training and feedback usually provided by a trained teacher. However, this type of training is expensive, requires a substantial time-commitment. The automated pronunciation error detection method, which identifies the erroneous phonemes from continuous speech, will allow L2 learners to work economically and efficiently, and improve the efficacy of current teaching methods.

Many automated error detection methods have been developed using ASR-based confidence scores [1, 2]. This approach has an advantage in the easy implementation; the score can be obtained easily from ASR system. However, it has disadvantage in the specialization for the specific phonemes with which L2 learners make frequent errors. In the beginning stage, L2 learners tend to make pronunciation errors on L2 phonemes which do not exist in their native language (L1), and some of these errors may remain even after several years of learning. The pronunciation training methods need to take into special consideration for these phonemes, but it is difficult for the confidence score based method since the scores are calculated for all phonemes in a similar way.

We developed a method which is a combination of the confidence scoring and Landmark-based SVMs. The potential errors were predicted based on the L1/L2 phonology and ESL literature, and the landmark-based SVMs were trained for them. Finally, the landmark-based SVMs were combined with confidence scoring methods. The method was tested on L2 learners' spontaneous speech described in [3].

This paper will proceed as follows: we will review previous studies (section 2), present the structure of the method (section 3), and report experiment setup (section 4). The results will be presented (in section 5), and compared with the previous studies in depth (in section 6).

2. Previous studies

The confidence score based method has been frequently used in this field [1, 2]. The Goodness of Pronunciation measure in [2] measures how closely each phone in the utterance matches the recognizer's acoustic model. Mismatches result in low scores, which provide a profile of the speakers' production errors.

Recently, researchers have investigated the use of classifiers in automated pronunciation scoring [4–7] and showed that the classifier method is more effective in implementing targeted phoneme-specific scoring.

Troung et al. [4] and Strik et al. [5] developed the acoustic-phonetic-feature-based classifier (AP-classifier) and cepstral-coefficients-based classifier (MFCC-classifier) in Dutch /x/ error detection. Doremalen et al. [8] extended this approach to 11 Dutch vowels. They focused on phonemic substitution errors that L2 learners mistakenly replaced an L2 phoneme with a different phoneme. In both studies, the classifier method achieved higher accuracy than confidence score.

In [4, 5, 8], only the features near the stop release were selected. This approach, especially MFCC-classifier, is closely related with landmark-based SVMs [9]. A landmark is a sudden signal change, and stop release is a landmark. Landmark-based SVMs, which were trained only using the spectral features extracted from the frame including and adjacent to a landmark, achieve high accuracy in the binary distinctive feature classification (e.g. distinction between stop and fricative consonants, high and low vowels.)

Yoon et al. [10] applied landmark-based SVMs systematically in error detection of 8 phonemes. Landmark-based SVMs is easy to apply to the various consonants and vowels, since it does not require the implementation of a phoneme specific feature extraction algorithm. The method was tested on artificial L1 data in which pronunciation errors were simulated by redefining the pronunciation of particular words. It achieved a promising result in this data; the method achieved a comparable performance to the confidence scoring, and the combination of the two methods with development test data could achieve further improvements.

In this study, the method was tested on speech collected from English learners with intermediate proficiency. In many previous studies, test data were speech from low proficiency L2 learners or artificial L1 speech, and most errors were phonemic: substitution of two different phonemes, insertion/deletion of a phoneme. In contrast, a high proportion of the errors in this data were distortion errors. Distortion is an error which cannot be classified into insertion, deletion or substitution. The non-categorical substitution, which is neither target-like nor clearly a substitution, is one example of 'distortion'. For instance, differences in voice onset times (VOT) for voiceless stop consonants were designated as 'distortion' when the VOT values were too short or too long for the categorical placement of the targeted phoneme, but not different enough to nudge the production into a different category. The high proportion of the distortion errors may increase the difficulty of the error detection.

3. Method

3.1. Overview

From the ESL literature, English phonemes with which L2 English learners make frequent errors were selected, and SVMs were trained in order to distinguish the errors from the correct phones. The landmark-based SVM method was combined with the confidence scoring method. A score-combination SVM was trained using the development test data. In the test, confidence score and landmark-based SVM score were calculated for each phone and combined using the score-combination SVM. If the SVM score was lower than a phoneme specific threshold, the phone was classified as an error.

3.2. Confidence Score

The speech was aligned against the manual transcription using a speech recognizer, and the target L2 phonemes were automatically extracted from this time-aligned phoneme segmentations. The confidence score was calculated using the acoustic model of the speech recognizer. For each phone, the frame-based confidence score was calculated as in [2].

3.3. SVMs

For each phoneme selected from the ESL literature, one SVM was trained in order to distinguish the target phoneme from the substitution. For each pair, the target phoneme was the positive example, while the possible substitution phone was the negative example. For example, if the target English phoneme was [f], and its potential substitution pattern was [p], then [f] was classified as a positive example, while [p] was classified as a negative example, and an SVM classifier was trained in order to distinguish [f] from [p]. For each pair, the same numbers of positive examples and negative examples were used for training.

All SVMs in this study are based on the acoustic feature vector including 39 PLPs (12 PLP coefficients, energy, their deltas and acceleration, computed once/10ms with a 25ms window) and formants (F1 and F2) extracted from [11]).

For vowels, 3 frames from the middle point were selected, and all feature vectors were concatenated (41x3). For consonants, 3 frames each from the initial, middle, and final points were selected and all feature vectors were concatenated (41x9). The frames were selected based on landmark theory and [9].

4. Experiment

4.1. L2 phoneme selection

In this study, the method was implemented for Korean learners. 6 phonemes (hereafter, 'difficult phones') with which Korean speakers make frequent pronunciation errors were selected from [12]. For each phoneme, its potential substitution error pattern was collected from [12].

Table 1 provides 6 pairs of L2 target phonemes and their possible substitutions. All symbols used in pronunciation columns are International Phonetic Alphabet.

Table 1: *Target English phonemes and their potential substitution patterns by Korean*

L2 phon.	Subst. phon.	Original word	Original pronunciation	Subst. pronunciation
æ	e	cap	k ae p	k e p
ɪ	i	bit	b ɪ t	b i t
l	ɾ	light	l aɪ t	ɾ aɪ t
θ	s	thick	θ ɪ k	s i k
v	b	vase	v eɪ s	b eɪ s
ð	d	they	ð eɪ	d eɪ

4.2. Data

Four different sources of data were used in the training. Table 2 presents the size and the source of the training and test data.

Table 2: *Training and test data*

		Size (hours)	Num. of speakers	Corpus
Train	Acoustic model	50	1953	HUB4
	Landmark SVMs	2	450	TIMIT
	Score combination	0.7	15	Buckeye
Test		0.5	5	Rated Speech Corpus

All training data are L1 data, while test data are L2 data. For the development of the automated pronunciation error detection, L2 learners' speech data, where the accuracy of each phone was rated, is required. The Rated speech corpus of L2 English learners [3] was used in the evaluation. The phone accuracy rating and the distribution of errors will be reported in detail in 4.7.

4.3. Acoustic model training for confidence score

A stress/gender-dependent triphone model was trained on the 1997 HUB4 English data [13] using the HTK toolkit [14]. From HUB4 data, English broadcast news, about 50 hours of sound files spoken by native English speakers were used in the training. The best phone accuracy was achieved by the model with 13 Gaussian mixtures. The phone accuracy rate in HUB4 evaluation data was 61%.

4.4. SVMs

SVMs were trained using TIMIT data (a wideband read speech corpus)[1]. Among the 6300 sentences in TIMIT, only the phonetically compact 'sx' sentences were selected. A total of 2310 sentences from 450 speakers were used for training. SVMs were trained using a Radial Basis Function (RBF) kernel using the SVM-light toolkit [15].

4.5. Score Combination

Score combination SVMs were trained using Buckeye Corpus of conversational speech [16]. A linear-kernel-based SVM was trained using the SVM-light toolkit with confidence score, SVM score, and phoneme id as input features.

4.6. Phoneme-Specific Threshold

Witt [2] pointed out the range of scores differs according to the phoneme. For instance, the scores of fricative consonants have a broader distribution than vowels. She showed that using different thresholds for each phoneme results in an improvement in the accuracy of the error detection.

In [10], phoneme-specific thresholds were found using the development test data which are from the same corpus with test data. In the current study, all L2 data were used in the evaluation due to small data size, and no development test data were available. Due to this problem, the mean score of each phoneme was calculated from the Buckeye Corpus, and used as a phoneme-specific threshold. The thresholds were found for confidence score, SVM score, and combined score separately.

4.7. Phone errors in L2 data

The Rated speech corpus of L2 English learners contained 28 L2 speakers speech representing 6 language backgrounds. For each speaker, approximately 6.5 minutes of spontaneous speech were collected.

Phone accuracy rating is costly and time consuming work. Currently only 30% of data (13 speakers' speech) have been rated. Among 13 rated speakers, 5 Korean speakers' speech were used in this study. All 5 Korean speakers were intermediate students. Detailed information is provided in [10].

Two phoneticians with intensive ESL teaching experience assigned the phone accuracy scores. Each phone was labeled using a binary score ('correct' or 'error'). Inter-rater reliability was 89%, while intra-rater reliability of the two raters were 96% and 92%.

The error category was further classified as 'substitution', 'insertion', 'deletion' or 'distortion'. Table 3 presents the proportions of subcategories in total errors from two raters.

Distortion was the most frequent sub-category, followed by substitution. Distortion and substitution occupied approximately 75% of total errors. The high proportion of the 'distortion' class suggested that L2 learners in this study made non-categorical substitution most frequently.

Table 3: Distribution of error sub-categories

	Subs.	Del.	Ins.	Distortion
Proportion in total errors (%)	29.9	14.9	11.0	44.2

Intermediate learners may make fewer errors compared to the beginner learners who were recruited for the previous studies such as [2]. In fact, the error ratio, which is the proportion of error phones in the total phones, was 7.78 % in average[2]. The low proportion of the errors made the evaluation of the method more difficult. In order to measure the impact of difficult phones on total errors, the ratio of difficult phone errors was calculated by counting the number of errors involving a difficult phone divided by the number of total errors. The ratio was 40%.

5. Results

The performance of the algorithm was evaluated using an F-score measure. Table 4 presents the F-scores of each method on the test data. Due to the low proportion of errors, the test data were adjusted to include same numbers of the correct samples and errors; the same numbers of correct phones were randomly selected from L2 data. The majority class baseline is 0.50 for all phonemes.

Table 4: *F-scores for each phoneme*

F-score	æ	ɪ	θ	ð	v	l	mean
Confidence	0.63	0.52	0.66	0.57	0.54	0.69	0.60
SVM	0.73	0.49	0.60	0.65	0.78	0.78	0.67
Combined	0.69	0.48	0.54	0.63	0.72	0.80	0.64

The confidence score shows higher F-scores for [ɪ, θ], while SVM score shows higher F-scores for [æ, ð,v,l]. Average F-score of the landmark-based SVM was 0.67, while that of the confidence-scoring system was only 0.60.

The combined method did not achieve further improvement. It was approximately 3% lower than landmark-based SVM method in absolute value.

6. Discussion

Table 5 provides the comparison of results between [10] and this study. This study replicated the methods of [7] on two different evaluation data; in [10], the method was tested on native English speakers' speech (hereafter, L1 data) in which the pronunciation errors were simulated by redefining the pronunciation of the particular words. For instance, rescoring software was told that the word 'pilot' contains [f], but the original speech remained unchanged. Thus, the data included artificial pronunciation errors which imitate the patterns of L2 learners.

The method achieved an F-score of 0.67 in L2 data and an F-score of 0.85 in the artificial L1 data; there was about 18% decrease in F-score in absolute value. The decrease of F-score

[1] Hasegawa-Johnson et al. [9] showed that the accuracy of landmark based SVMs decreased significantly when the training and test data were from different corpus. Since test data are laboratory speech without background noise, TIMIT data were used instead of Broadcast news data.

[2] The error ratio ranged from 3.76 % to 10.43 %.

Table 5: *Comparison between [10] and the current study*

Data	Confidence	SVM	Combined
Artificial L1 data ([10])	0.83	0.81	0.85
L2 data (current study)	0.60	0.67	0.64

in real L2 data is predictable. L1 data contained only substitution errors (this being an inherent limitation of the method used to generate artificial errors in L1 data), whereas the real L2 data are dominated by harder-to-detect distortion errors.

All L2 speakers in this study were intermediate learners, and the proportion of the 'distortion' class was high. The less salient difference between the correct phones and errors may increase the difficulty of the error detection, and therefore F-scores decreased. However, the results are still inspiring; the landmark-based SVMs achieved superior accuracy to the confidence scoring method without any L2 training data. In addition, with small training data sizes, SVMs have an advantage over the confidence scoring method. SVM training data is thus 25 times smaller than the acoustic model training data.

The high proportion of 'difficult phones' on total errors strongly supports the appropriateness of the current approach which predicts the potential errors based on L1/L2 phonology first, and enhances the method for them. L1 phonology influences L2 pronunciation not only for beginners but also intermediate learners. Therefore, a method specialized for these phonemes will be beneficial for both beginners and intermediate learners.

The combination of two methods did not improve the accuracy in L2 data. This result is different from [10]'s results where the combination of two methods lead to a statistically significant improvement. This suggests the importance of the appropriate development test data; in both studies, L1 development test data were used to select thresholds and stream weights. The training of score combination in the same L2 data may result in the additional improvement.

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 valuable feedback to correct the error. Based on this information, accurate feedback on how to correct the pronunciation error can be provided to L2 learners.

7. References

[1] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, "Automatic scoring of pronunciation quality," in *Speech Communication*, 2000, pp. 88–93.

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

[3] S. Yoon, L. Pierce, A. Huensch, E. Juul, S. Perkins, R. Sproat, and M. Hasegawa-Johnson, "Construction of a rated speech corpus of l2 learners' speech," *CALICO*.

[4] K. Truong, A. Neri, C. Cuchiarini, and H. Strik, "Automatic pronunciation error detection: an acoustic-phonetic approach," in *the InSTIL/ICALL Symposium*, 2004, pp. 135–138.

[5] H. Strik, K. Truong, F. de Wet, and C. Cucchiarini, "Comparing classifiers for pronunciation error detection," in *Proceedings of Interspeech 07*, 2007, pp. 1837–1840.

[6] F. Pan, Q. Zhao, and Y. Yan, "Mandarin vowel pronunciation quality evaluation by a novel formant classification method and its combination with traditional algorithms," in *Proceedings of ICASSP 08*, 2008, pp. 5061–5064.

[7] A. Ito, Y.-L. Lim, M. Suzuki, and S. Makino, "Pronunciaiton error detection method based on erro rule clustering using a decision tree," in *Proceedings of interspeech 05*, 2005, pp. 173–176.

[8] J. van Doremalen, C. Cucchiarini, and H. Strik, "Automatic detection of vowel pronunciation errors using multiple information sources," in *In Proceedings of ASRU 2009*, 2009.

[9] M. Hasegawa-Johnson, J. Baker, S. Borys, K. Chen, E.Coogan, S. Greenberg, A. Juneja, K. Kirchhoff, K. Livescu, S. Mohan, J. Muller, K. Sonmez, and T. Wang, "Landmark-based speech recognition: Report of the 2004 johns hopkins summer workshop," in *Automatic Speech Recognition and Understanding Workshop*. ICASSP, 2005.

[10] S. Yoon, M. Hasegawa-Johnson, and R. Sproat, "Automated pronunciation scoring using confidence scoring and landmark-based SVM," in *In Proceedings of InterSpeech 2009*, 2009.

[11] P. Boersma and D. Weenink, *Praat: doing phonetics by computer (Version 4.5.02) [Computer program]*, 2006.

[12] M. Swan and B. Smith, *Learner English*. Cambridge: Cambridge University Press, 2002.

[13] D. Pallet, "Overview of the 1997 darpa speech recognition workshop," in *DARPA Speech Recognition Workshop*. DARPA, 1997.

[14] S.Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version3.2)*. Microsoft Corporation and Cambridge University Engineering Department, 2002.

[15] T. Joachims, *Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning*, B. Scholkopf, C. Burges, and A. Smola, Eds. MIT-Press, 1999.

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