Landmark-based Automated Pronunciation Error Detection

Su-Youn Yoon[1], Mark Hasegawa-Johnson[2], and Richard Sproat[3]

[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 at the phone level and landmark-based Support Vector Machines (SVMs). Landmark-based SVMs were implemented to focus the method on targeting specific phonemes in which L2 learners make frequent errors.

The method was trained on the phonemes that are difficult for Korean learners and tested on intermediate Korean learners. In the data where non-phonemic errors occurred in a high proportion, the SVM method achieved a significantly higher F-score (0.67) than confidence scoring (0.60). However, the combination of the 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, a method that is specialized for these phonemes would be beneficial for both beginners and intermediate learners.

Index Terms: automated pronunciation error detection, computer-aided pronunciation training systems, phone-level confidence scores, 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 diagnosis followed by training and feedback that is usually provided by a trained teacher. However, this type of training is expensive and 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 phone-level confidence scores [1, 2]. This approach has an advantage of easy implementation - the score can be easily obtained from an ASR system. Furthermore, it can be applied to all L2 learners and is not limited by their native language (L1) background. However, it has the disadvantage in its specialization for the specific phonemes in which L2 learners make frequent errors. In the beginning stage of learning a new language, L2 learners tend to make pronunciation errors on L2 phonemes which do not exist in their L1, and some of these errors may remain, even after several years of learning. The pronunciation training methods need to incorporate special considerations for these phonemes, but it is difficult because the scores are calculated for all phonemes in a similar way.

We developed a method which is a combination of the phone-level confidence scoring and Landmark-based SVMs[1] 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 phone-level confidence scoring methods and were tested on L2 learners' spontaneous speech, as 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 the 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 (GOP) in [2] measures how closely each phone in an 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 an acoustic-phonetic feature-based classifier (AP-classifier) and a cepstral-coefficient-based classifier (MFCC-classifier) for Dutch /x/ error detection. Doremalen et al. [8] extended this approach to 11 Dutch vowels. They focused on phonemic substitution errors where L2 learners mistakenly replaced an L2 phoneme with a different phoneme. In both studies, the classifier method achieved higher accuracy than the GOP score. In [4, 5, 8], only the features near the stop release were selected. This approach, especially the MFCC-classifier, is closely related to landmark-based SVMs [9].

Yoon et al. [10] applied landmark-based SVMs systematically in the error detection of 8 phonemes. The method was tested on artificial L1 data in which pronunciation errors were simulated by redefining the pronunciation of particular

[1]A landmark is a sudden signal change. For example, a 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, achieved high accuracy in the binary distinctive feature classifications (such as the distinction between stop and fricative consonants, and high and low vowels.)

words. A promising result was achieved in this data; the method achieved a comparable performance to the GOP-based method, 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 contained speech from low proficiency L2 learners or artificial L1 speech, and most errors were phonemic (e.g., the substitution of two different phonemes, insertion/deletion of a phoneme). In contrast, a high proportion of the errors in our data were distortion errors. Distortion is an error which cannot be classified as an 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 distortion errors may increase the difficulty of error detection.

3. Method

3.1. Overview

From the ESL literature, English phonemes in 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 GOP-based method. A score-combination SVM was trained using the development test data. In the test, a GOP score and landmark-based SVM score were calculated for each phone and then 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. GOP score

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

3.3. SVMs

For each phoneme selected from the ESL literature, one SVM was trained in order to distinguish the targeted phoneme from the substitution. For each pair, the targeted phoneme was the positive example, while the possible substitution phone was the negative example. For example, if the targeted 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 the two. 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 perceptual linear prediction (PLP) (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, this method was implemented for Korean learners of English. 6 phonemes (hereafter, 'difficult phones') in 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 from the International Phonetic Alphabet.

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

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 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 were L1 data, while test data were L2 data. For the development of the automated pronunciation error detection, L2 learners' speech data, where the accuracy of each phone was rated, was 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 broadband read speech corpus)[2]. 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 the Buckeye Corpus of conversational speech [16]. A linear-kernel-based SVM was trained using the SVM-light toolkit with GOP score, SVM score, and phoneme id as input features.

4.6. Phoneme-Specific Threshold

Witt [2] pointed out that 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 the improved accuracy of the error detection.

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

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 the data (13 speakers' speech) have been rated. Out of the 13 rated speakers, this study used speech from 5 Korean speakers. 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'). The inter-rater reliability was 89%, while the 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 proportion of subcategories in the total errors from two raters.

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

Table 3: Distribution of error sub-categories

	Subs.	Deletion	Insertion	Distortion
Proportion in total errors (%)	29.9	14.9	11.0	44.2

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

Intermediate learners may have made 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 on average 7.78 %[3]. The low proportion of 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 number of correct samples and errors; the same number 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
GOP	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 GOP shows higher F-scores for [ɪ, θ], while the SVM score shows higher F-scores for [æ, ð, v, l]. The average F-score of the landmark-based SVM was 0.67, while that of the GOP system was 0.60.

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

6. Discussion

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

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

remained unchanged. Thus, the data included artificial pronunciation errors which imitated the patterns of L2 learners.

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

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

The method achieved an F-score of 0.67 in the L2 data and an F-score of 0.85 in the artificial L1 data; there was approximately 18% decrease in F-score. The decrease in the F-score in real L2 data is predictable. The L1 data contained only substitution errors (this being an inherent limitation of the method used to generate the artificial errors in the L1 data), whereas the real L2 data were dominated by harder-to-detect distortion errors.

All L2 speakers in this study were intermediate learners, and the proportion of the 'distortion' class errors was high. The less salient difference between the correct phones and errors may have increased the difficulty of 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 the two methods did not improve the accuracy in the L2 data. This result is different from Yoon et al. [10]'s results where the combination of two methods led 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 the score combinations in the same L2 data may have resulted in additional improvements.

7. Conclusion

In this study, we developed a pronunciation error detection method based on a GOP score and a landmark-based SVM score. Landmark-based SVMs were specialized for the specific phonemes in which L2 learners make frequent errors, and it achieved a superior performance over the GOP score in those selected phonemes. However, the performance of the landmark-based SVMs will be influenced by the prediction of the 'difficult phonemes'. If the speakers make pronunciation errors on the phoneme other than the predicted phonemes, or the error phones are different from the predicted patterns, the landmark-based SVMs may not achieve good performance. Furthermore, since the 'difficult phone' and the error pattern are heavily influenced by their L1 background, the performance will be influenced by learner's L1. On the contrary, the GOP score is not influenced by error prediction, and it can be applied to L2 learners without the limitation of their L1 background.

Therefore, the two methods are complementary, and the combination of the two methods may be beneficial. For instance, the GOP scores can be used as the initial test which can be applied to all learners, and landmark-based SVMs can be used as the extended test, specialized according to learner's L1. In addition, SVM can be used effectively in the pronunciation training; for instance, if a speaker cannot distinguish [f] from [p], the SVMs, trained on the [p,f] pair, can give feedback whether the sound is a [p] or an [f]. This information can be used as a key to provide valuable feedback to train the L2 phoneme.

8. 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.