ASSESSMENT OF NON-NATIVE SPEECH USING VOWEL SPACE CHARACTERISTICS

Lei Chen and Keelan Evanini

Educational Testing Service
Princeton, NJ

Xie Sun

University of Missouri
Columbia, MO

ABSTRACT

In this paper, we propose the idea of using the characteristics of a speaker's vowel space for automated assessment of second language (L2) proficiency. Specifically, we adpot features that were shown in previous studies to be good indicators of native speaker intelligibility and clarity and apply them to L2 speech from non-native speakers. The features focus on three peripheral vowels (IY, AA, and OW) and measure a speaker's coverage of the vowel space. A pilot study and a large-scale corpus study involving read speech produced by native and non-native speakers were conducted in which the vowel space features were rank correlated with pronunciation scores provided by human listeners for the non-native speech and an assumed higher score for the native speech. The results of the studies show that several of the features achieve moderately high correlations with the pronunciation scores, supporting their usefulness for automated assessment of non-native speech. The feature with the best performance in the large-scale study was the $\mathrm{F2} - \mathrm{F1}$ distance for IY, which achieved a correlation of 0.78 with pronunciation proficiency scores.

Index Terms: speech assessment, phonetics, vowel quality

1. INTRODUCTION

In the last two decades, there have been many studies on using automatic speech recognition (ASR) technology to assess non-native read speech [1, 2, 3]. In these previous studies, features calculated from the recognition or forced-alignment process (e.g., average ASR confidence values, speaking rate, etc.) were widely used as features for predicting human speaking proficiency. Features derived from acoustic phonetic measurements, however, have rarely been implemented. In the case of vowels, duration has been utilized as an indicator for fluency [4], but features using formant measurements of vowel quality have not been sufficiently studied.

Neglecting the fine-grained acoustic information of vowel spaces in automated speech assessment causes several problems. First, important perceptual cues used by humans to judge pronunciation are not utilized. Second, a large amount of previous research in phonetics, second-language acquisition, and L2 instruction can not be utilized. Third, most L2 teachers are much more familiar with concepts related to vowel quality than the engineering-oriented confidence measurements used by the current automated speech assessment systems; using features related to vowel quality may make automated speech assessment more easily understood by teachers. Finally, detailed information about a speaker's vowel space can help the assessment system provide important feedback to test-takers regarding the required adjustments to their pronunciation to sound more native-like. Therefore, in this paper, we report on two studies analyzing vowel space characteristics and their relationship to pronunciation scores provided by human annotators.

This paper is organized as follows: Section 2 reviews previous related research; Section 3 presents definitions of the vowel features used in this study and describes the procedure used to obtain vowel formant measurements; Section 4 reports on a small pilot study examining the relationship between vowel space characteristics and pronunciation scores; Section 5 presents the results of a large-scale corpus study which employed similar methodology; and Section 6 discusses our findings and proposes future research directions.

2. PREVIOUS RESEARCH

Several previous studies have examined vowel space characteristics in relation to a native speaker's perceived intelligibility or clarity of speech. The main idea in these studies is that clearly articulated, intelligible speech is characterized by less vowel reduction and, thus, more extreme formant measurements for vowels on the periphery of the vowel space. In one influential study, [5] compared the intelligibility of native speakers with several vowel space features derived from the three vowels IY (as in 'seek'), AA (as in 'sock'), and OW (as in 'soak'). Their results showed that speakers with larger vowel spaces were generally rated as more intelligible.

[6] conducted an acoustic study of real and imagined foreigner-directed speech by native speakers of English. They also extracted several features characterizing a vowel space's expansion, and found that the vowel space was more expanded for foreigner-directed speech, i.e., in situations when the speakers are attempting to increase the intelligibility of their speech. In a similar study, [7] found that vowel formants were more peripheral in clear speech as opposed to conversational speech. In a study that investigated native speech in a

language other than English, [8] show that native speakers of Cantonese also have expanded vowel spaces when producing clearer speech.

Several phonetics studies have also investigated vowel space measurements of non-native speakers in relation to L2 proficiency. For example, [9] investigated the effect of L2 experience on non-native speakers' production and perception of pairs of English vowels that are easily confused. They computed features based on differences in the F1 and F2 dimensions, and found that more experienced speakers produced vowels that were closer to the native speaker productions. In another study, [10] explored the effect of phonetic speech training on vowel space characteristics in Croatian. By comparing vowel spaces occurring in Croatian produced by native-Croatian speakers and two groups of non-native speakers (English and Spanish) before and after phonetic training, significant improvements in the vowel space characteristics were shown.

Finally, vowel space analysis has also been used as an aid in computer-assisted pronunciation training. For example, [11] designed software to let users control a virtual ball on a computer screen to match native speakers' vowel space locations. Their system was found to help language learners to better control the vowel space of a new language.

Clearly, acoustic analyses of vowel spaces have been useful in studies of native speaker intelligibility and non-native speaker L2 proficiency. However, no studies to date have specifically compared vowel space features in L2 speech with human ratings of pronunciation quality. This paper attempts to address that gap and directly explore the utility of vowel space features in automated L2 proficiency assessment.

3. VOWEL SPACE FEATURES

Several measurements have been previously proposed to describe the characteristics of vowel spaces among different speakers. The basic idea is that the use of more peripheral vowels is critical to good pronunciation, thus causing higher intelligibility and a perception of nativeness. In this study, five features related to vowel space expansion that were suggested by [5] to be relevant for native speaker intelligibility will be explored in this paper in the context of L2 proficiency assessment: range, area, overall dispersion, within-category dispersion, and F2–F1 distance.

For all of the features used in this paper, the three peripheral vowels IY, AA, and OW are investigated in detail. These three vowels generally represent the most extreme points in a speaker's vowel space, and are thus most useful for determining overall characteristics about a speaker's coverage of the vowel space.[1]

[1] As was also done by [5], the vowel UW (as in 'goose') was not used as the peripheral vowel in the lower F2 range, due to its widespread fronting in many dialects of English. All experiments described below were also conducted with UW instead of OW, but the results were worse, suggesting that OW is a better representitave of the back periphery of the vowel system.

3.1. Vowel space range

The vowel space range represents the simplest method of determining a speaker's coverage of the vowel space. It is calculated by subtracting the overall minimum value from the maximum value for both F1 and F2. As described above, this feature used the three peripheral vowels IY, AA, and OW; due to the nature of the distributions for these vowels, the range features were in most cases identical to: $F1Range = Max_{F1}(AA) - Min_{F1}(IY)$ and $F2Range = Max_{F2}(IY) - Min_{F2}(OW)$. Since it relies on formant measurements from individual vowel tokens, this feature can be sensitive to outliers.

3.2. Vowel space area

The area of the vowel triangle defined by the mean F1 and F2 values of the three peripheral vowels was used as a measure of the overall coverage of the vowel space. This feature was calculated using Heron's formula for the area of a triangle:

$$area = \sqrt{s(s - D_{\bar{IY},\bar{AA}})(s - D_{\bar{AA},\bar{OW}})(s - D_{\bar{OW},\bar{IY}})}$$

where $s = 0.5 \times (D_{\bar{IY},\bar{AA}} + D_{\bar{AA},\bar{OW}} + D_{\bar{OW},\bar{IY}})$, $\bar{V}$ represents the mean F1 and F2 values for vowel V, and $D_{x,y}$ represents the Euclidean distance between two values in the F1-F2 plane:

$$D_{x,y} = \sqrt{(\mathrm{F1}_x - \mathrm{F1}_y)^2 + (\mathrm{F2}_x - \mathrm{F2}_y)^2}$$

3.3. Vowel space dispersion

Additionally, the vowel space dispersion, defined as the average distance from individual peripheral vowel tokens to the center of the vowel space, was used. The $F1$ and $F2$ values of the vowel space center, $\bar{V}$, are the overall mean values of $F1$ and $F2$ computed using all of a speaker's vowel tokens from all vowel categories. Thus, the vowel space dispersion is defined as:

$$dispersion = \frac{\sum D_{IY_i,\bar{V}} + \sum D_{AA_i,\bar{V}} + \sum D_{OW_i,\bar{V}}}{N}$$

where N is the total number of tokens of vowels in the categories IY, AA, and OW.

3.4. Within-category vowel dispersion

The within-category dispersion measures how far the tokens for each of the three peripheral vowels (IY, AA, and OW) are from the their respective category mean values (as opposed to measuring the distance between the individual vowel tokens and the overall mean F1 and F2 values, as was done

for the overall vowel dispersion feature). Thus, this feature shows how spread apart the tokens of each of the three vowels classes are. The equation for within-category vowel dispersion is as follow:

$$\frac{1}{3} * (\frac{\sum D_{IY_i,\bar{IY}}}{N_{IY}} + \frac{\sum D_{AA_i,\bar{AA}}}{N_{AA}} + \frac{\sum D_{OW_i,\bar{OW}}}{N_{OW}})$$

3.5. $\mathrm{F2} - \mathrm{F1}$ distance

Finally, the F2–F1 values for the peripheral vowels IY and AA were suggested to be effective for measuring the extreme points in the vowel space. Among all vowels, the F2–F1 distance is generally largest for IY and smallest for AA; thus, [5] hypothesized that the F2–F1 distance would be positively correlated with intelligibility for IY (greater distances mean more peripheral tokens of IY) and negatively correlated with intelligibility for AA (smaller distances mean more peripheral tokens of AA). For each speaker, we thus calculated an overall F2–F1 score for IY and AA by taking the average F2–F1 score for each of the speaker's tokens of the relevant vowel.

3.6. Vowel formant measurements

All vowel formant measurements used to compute the features described in this section were extracted automatically from the speech signal using the following procedure. First, each spoken response was aligned with text of the prompt using the Penn Phonetics Forced Aligner (P2FA) [12] to determine word- and phoneme-level boundaries. Next, Praat [13] was used to extract F1 and F2 measurements at the point 1/3 of the way into the duration of the vowel.[2] Only vowels bearing lexical stress (in the CMU dictionary) were included in the feature computation, since unstressed vowels are expected to be substantially centralized, and would thus affect the feature values. Additionally, all vowel tokens preceding the consonant R were excluded from the analysis, due to the strong centralizing effect that R has on preceding vowels. Furthermore, to exclude outliers, all tokens that were greater than 3 standard deviations away from the mean value for that vowel category were omitted. Finally, the vowel tokens for each speaker were normalized to reduce effects of speaker-specific physiological characteristics by taking z-scores of all of the formant measurements.

4. PILOT STUDY

This section describes a small pilot study that was conducted to explore the potential usefulness of the features described in Section 3 for the purpose of English proficiency assessment.

[2]Praat's default LPC method of formant prediction based on was used with 5 predicted formants and a maximum formant value of 5000 Hz.

Vowel	Word Tokens
IY	*each*, *needs*, *week*
AA	*projects*, *quality*, *want*
OW	*located*, *open*, *over*

Table 1. Words used in the Pilot Study

4.1. Data and Methodology

For this pilot study, a single Read Aloud item was selected for analysis from among several responses provided by speakers in an English proficiency assessment. This item consists of a paragraph containing 96 words which the speakers were instructed to read out loud in a natural manner. The entire response was then scored by experienced human raters using a three-point scale for overall pronunciation assessment.

For each of the three score levels, we selected 5 female and 5 male speakers who all shared the same L1 for analysis. In addition, the same paragraph was read by two female and two male native speakers of American English. As a result, this study contains speech data corresponding to four score levels from a total of 34 speakers: low-level (NNS1), medium-level (NNS2), and high-level (NNS3) for Non-Native Speakers, as well as Native Speakers (NS).[3]

All stressed tokens of the peripheral vowels IY, AA, and OW were used to compute the vowel space features, subject to the exclusions described in Section 3.6. The relevant words contained in the Read Aloud item from this assessment are listed in Table 1.

4.2. Results

To determine the usefulness of a feature to discriminate among the different pronunciation proficiency levels, we calculated the Spearman rank order correlation coefficient, ρ, between the pronunciation scores and each of the vowel space features. These results are summarized in Table 2.

Feature	Spearman correlation
F1 Range	-0.08
F2 Range	0.11
Area	0.01
Dispersion	0.31
Within-category dispersion	-0.15
$\mathrm{F2} - \mathrm{F1}$ **for IY**	**0.38**
$\mathrm{F2} - \mathrm{F1}$ **for AA**	**-0.47**

Table 2. Pilot Study results (features with correlations significant at $\alpha = 0.05$ are in bold)

[3]For the purpose of calculating the rank-order correlations below, it is assumed that all of the native speakers have a higher pronunciation proficiency level than all of the non-native speakers, even though no explicit score was provided for the native speaker responses.

As Table 2 shows, two of the vowel space features had significant correlations with pronunciation scores for the Read Aloud items from these 34 speakers: the $F2 - F1$ distance for IY and AA. The correlations were in the directions expected by the hypothesis that more peripheral vowels lead to more intelligible pronunciation.

The fact that the other vowel space features did not show a significant correlation with pronunciation scores in this pilot study is likely due to the fact that the number of tokens of the relevant vowels produced by each speaker was very limited. As Table 1 shows, the number of tokens in each vowel category for each speaker was three. With only such a small number of tokens per vowel category, it is difficult to obtain an accurate characterization of each speaker's vowel space. However, the results of the pilot study were suggestive, and warranted further research on a larger data set.

5. LARGE-SCALE STUDY

The results from the pilot study described in Section 4 were promising, but were drawn from a very small data set, both in terms of the number of speakers and the number of vowel tokens per speaker. In order to address this limitation, and investigate the generalizability of the proposed vowel space features for assessment of non-native speech, a second study was conducted with a larger amount of data.

5.1. Data and Methodology

In this experiment, 325 non-native speakers who shared the same L1 responded to four Read Aloud items each in an English proficiency assessment. Due to the design of the assessment, there were three distinct sets of four Read Aloud items, meaning that the speakers did not all produce the same lexical items, as they did in the pilot study. However, the number of tokens produced in each vowel category by each speaker was much higher, thus facilitating the comparison among speakers who read different items. As in the pilot study, the responses were scored by human raters on a three-point scale for pronunciation proficiency. In this study, the responses were subsequently transcribed (to eliminate the effect of reading errors on the forced alignment procedure) and processed using the P2FA forced alignment system. Vowel formant measurements were again extracted according to the methodology described in Section 3.6.

The total number of tokens produced by each speaker that were used to calculate the vowel features varied, due to the different sets of Read Aloud items, and the fact that speakers did not always produce the text accurately. The mean number of tokens (and standard deviation) for each vowel produced by the 325 speakers in this experiment are as follows: 16.2 (5.0) for IY, 10.7 (3.8) for AA, and 9.0 (2.6) for OW.

Since no native speaker responses exist for the items used in this experiment, a source of native speaker vowel measurements from another domain was substituted. We used the Atlas of North American English corpus, which includes data from 437 speakers throughout North America [14]. Several speakers from every dialect region were included in the sample. Each speaker participated in an interview consisting of spontaneous speech and targeted elicitation of specific lexical items. Approximately 300 vowel formant measurements were extracted for each speaker and were manually verified. This corpus thus provides the most detailed sample of vowel formant variation among native speakers of English in North America. The mean number of vowel formant measurements (and standard deviation) for the three peripheral vowels from the speakers in this corpus are as follows: 12.5 (5.9) for IY, 27.6 (8.6) for AA, and 18.1 (7.7) for OW.

5.2. Results

As in the pilot study, the usefulness of each feature at discriminating among the levels of pronunciation proficiency is determined by calculating the Spearman rank-order coefficients between the feature values and the pronunciation proficiency scores. Since each non-native speaker responded to four Read Aloud items in the large-scale experiment, it is possible to compute both item-level and speaker-level correlations between the proficiency scores and the vowel space features (this was not possible for the pilot study, since only a single Read Aloud item was used).[4] For the speaker-level results, all of the vowel tokens from a single speaker were pooled together to compute the speaker-level vowel space features, and the four pronunciation scores for the different items were added together. These results are summarized in Table 3.[5]

Feature	Spearman correlation	
	Item-level	Speaker-level
F1 Range	0.55	0.32
F2 Range	0.55	0.25
Area	0.43	0.58
Dispersion	0.23	0.34
Within-category dispersion	-0.17	-0.71
$F2 - F1$ for IY	0.63	0.78
$F2 - F1$ for AA	-0.42	-0.58

Table 3. Large-scale experiment results (all correlations are significant at $\alpha = 0.01$)

Table 3 shows that the correlations between all vowel

[4]The total number of non-native speakers included in the speaker-level analysis was 229, since some speakers did not respond to all 4 items in the assessment.

[5]As described above, the native speaker vowel measurements in this experiment were drawn from a different domain, and, thus, do not have separate items associated with each native speaker. As was done in the item-level analysis, the native speakers were all assigned a proficiency score at a level higher than the non-native speakers for the purposes of computing the rank-order correlations.

space features and pronunciation proficiency scores were significant and moderately strong. In addition, the use of speaker-level scores generally improved the correlation values–the only two exceptions were the features involving ranges. The best-performing feature was the $\mathrm{F2}-\mathrm{F1}$ distance for the vowel IY, with a correlation of $\rho = 0.78$.

Furthermore, the correlations for each feature had the polarity expected given the hypothesis that an expanded vowel space leads to higher pronunciation proficiency scores. As in the pilot study, the $\mathrm{F2}-\mathrm{F1}$ distance for IY was positively correlated with pronunciation scores, and the $\mathrm{F2}-\mathrm{F1}$ distance for AA had a negative correlation. The two range features, the area feature, and the overall dispersion feature all have positive correlations with pronunciation scores, indicating that an expanded vowel space leads to a rater's perception that the speaker is more native-like. (In contrast, the area feature was found to have no significant correlation with intelligibility scores for native speakers in [5] and [6].) The fact that the correlations for the within-category dispersion feature are negative corresponds well with the hypothesis in [5] that more tightly clustered distributions for individual vowels lead to higher intelligibility since inter-category confusions would be less likely. While their study using native speakers found no significant correlation between within-category dispersion and intelligibility ratings, the results for this feature were significant at both the item-level and the speaker-level.

Figure 1 presents boxplots of the distributions for four of the features in the item-level analysis (the plots for the other three features also show similar patterns). In each case, the plots display a monotonic trend for the mean value of the feature from the lowest non-native proficiency level to the native speakers. While there is substantial overlap between the three non-native categories, the difference between the native speaker distribution and the three non-native speaker distributions for each vowel space feature is quite pronounced.

Finally, Table 4 presents a correlation matrix showing how the vowel space features pattern with respect to each other for the item-level analysis in this experiment. All of the pairs except one show significant correlations, but none of the correlations has a magnitude greater than 0.70.

6. DISCUSSION

In this paper, we proposed the idea of using features derived from an acoustic analysis of the vowel space in the automated assessment of non-native speech. The results from the large-scale study demonstrate the potential usefulness of this approach, since all vowel space features showed significant correlations with proficiency level.

As mentioned in Section 5.2, the largest difference among the different proficiency levels is between the native speakers and a group consisting of all of the three levels of non-native speakers. This indicates that even many of the highest-proficiency non-native speakers are still quite far from the native-speaker targets. Separate correlation analyses were conducted with the native speaker data excluded. In this case, the best-performing feature was vowel space area, with $\rho = 0.28$ for the speaker level analysis.

	F2 Range	Area	Dispersion	w.c. Dispersion	F2 − F1 for IY	F2 − F1 for AA
F1 Range	0.44	0.54	0.42	0.07	0.53	-0.56
F2 Range	–	0.58	0.47	0.10	0.70	-0.37
Area	–	–	0.58	-0.29	0.68	-0.53
Dispersion	–	–	–	*n.s.*	0.44	-0.54
w.c. dispersion	–	–	–	–	-0.29	0.24
F2 − F1 for IY	–	–	–	–	–	-0.49

Table 4. Correlation matrix for the 6 vowel space features using the item-level features

The results of this study demonstrate that automated assessment of non-native speech can be enriched by exploring areas outside of the standard features based on the ASR process. To further explore the utility of this approach, future studies will integrate the proposed vowel space features into an automated scoring engine containing many other types of features in order to observe the effect they have on the overall prediction of proficiency scores.

7. REFERENCES

[1] J. Mostow, S.F. Roth, A.G. Hauptmann, and M. Kane, "A prototype reading coach that listens," in *Proc. AAAI*, 1994, pp. 785–792.

[2] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, "Automatic Scoring of Pronunciation Quality," *Speech Communication*, vol. 30, pp. 83–93, 2000.

[3] S. M. Witt, *Use of Speech Recognition in Computer-assisted Language Learning*, Ph.D. thesis, University of Cambridge, 1999.

[4] L. Chen, K. Zechner, and X Xi, "Improved pronunciation features for construct-driven assessment of non-native spontaneous speech," in *NAACL-HLT*, 2009.

[5] A. R Bradlow, G. M Torretta, and D. B Pisoni, "Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics," *Speech Communication*, vol. 20, no. 3-4, pp. 255–272, 1996.

[6] R. Scarborough, O. Dmitrieva, L. Hall-Lew, Y. Zhao, and J. Brenier, "An acoustic study of real and imagined foreigner-directed speech," in *Proc. of ICPhS 2007*, 2007.

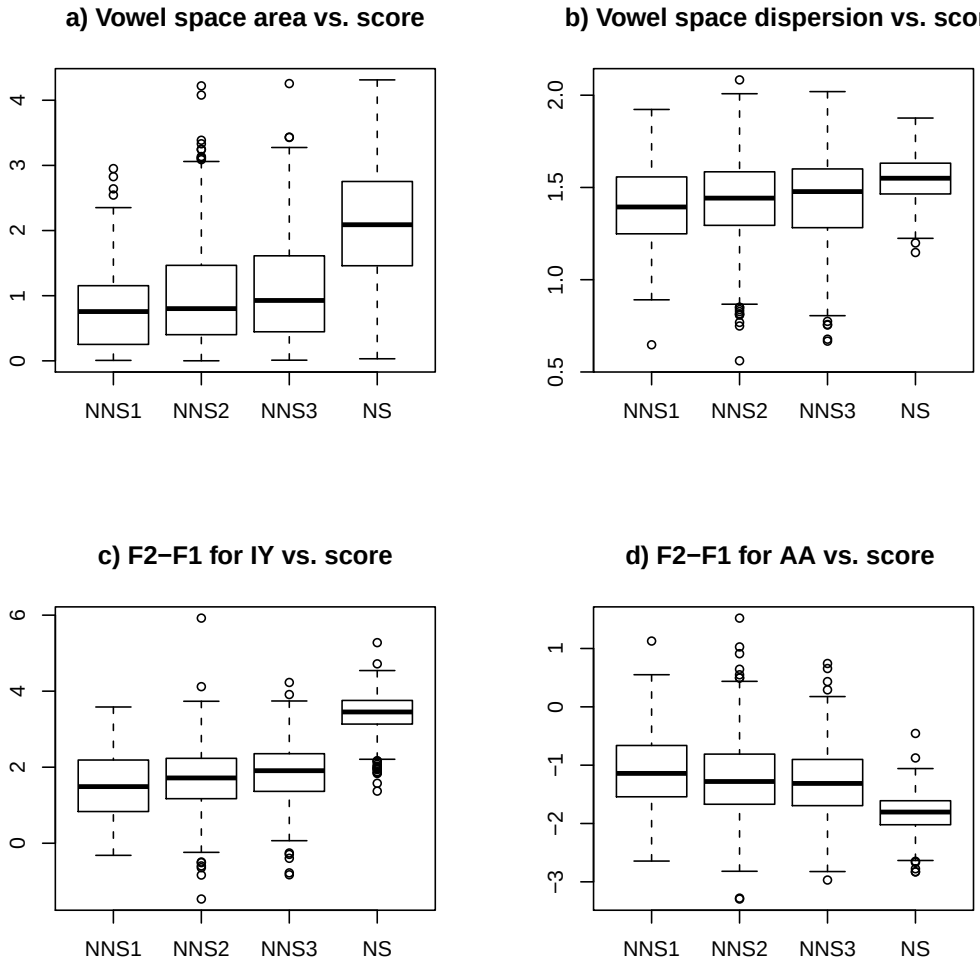

Fig. 1. Boxplots showing the relationship between vowel space features and pronunciation proficiency (item-level analysis)

[7] M.A. Picheny, N.I. Durlach, and L.D. Braida, "Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech," *Journal of Speech and Hearing Research*, vol. 29, pp. 434–446, 1986.

[8] H.C.N. Li and C.K. So, "Acoustic analysis of vowels spoken clearly and conversationally by non-native English speakers," in *Proceedings of the 11th Australian International Conference on Speech Science and Technology*, 2006.

[9] J. E. Flege, O. S. Bohn, and S. Jang, "Effects of experience on non-native speakers' production and perception of English vowels," *Journal of Phonetics*, vol. 25, no. 4, pp. 437–470, 1997.

[10] V. Mildner and D. Tomic, "Effects of phonetic speech training on the pronunciation of vowels in a foreign language," in *Proc. ICPhS*, 2007.

[11] P. Wik and D.L. Escribano, "Say 'Aaaaa': Interactive vowel practice for second language learning," in *Proc. SLaTE*, 2009.

[12] Jiahong Yuan and Mark Liberman, "Speaker identification on the SCOTUS corpus," in *Proc. of Acoustics '08*, 2008.

[13] P. Boersma and D. Weenick, "Praat: Doing phonetics by computer, version 5.0.38," *http://www.praat.org*, 2010.

[14] W. Labov, S. Ash, and C. Boberg, *The Atlas of North American English*, Mouton de Gruyter, 2006.