Does vowel quality really matter?

Yusuke Kondo[1], Aya Kitagawa[2], and Michiko Nakano[3]

[1]Open Education Center, [2]Graduate School of Education, [3]Faculty of Education and Integrated Arts and Sciences, Waseda University

yusukekondo@aoni.waseda.jp, aya@fuji.waseda.jp, nakanom@waseda.jp

Abstract

This paper reviews two of our researches on the relationship between vowel quality and overall evaluation in second language (L2) speech, and introduces another two researches on pronunciation by L2 learners in reading aloud. The results indicate that the pronunciation can be one of the predictors of the overall evaluation, but cannot be a dominant predictor of learners' proficiency.

Keywords

L2 speech, Reading aloud, Vowel quality

Introduction

In the previous researches (Kondo, Tsutsui, Nakano, Tsubaki, Nakamura, and Sagisaka, 2007; Kondo Tsutsui, Tsubaki, Nakamura, Sagisaka, and Nakano, 2007; Nakano, Kondo, and Tsutsui, 2008; Kondo and Nakano, 2009; and Kondo, Tsutsui, and Nakano, 2010), we have examined the relationship between speech characteristics and overall evaluation in an L2 reading aloud task. On the basis of the results of the researches, we constructed an automatic scoring system for L2 read-aloud speech. The system predicts examinees' score by using two speech characteristics, the indices of speech rate and rhythm, which are statistically significant predictors of the evaluation score given by human raters. Through these studies, we examined the pronunciation made by L2 learners of English and investigate the relationship between their characteristics in the pronunciation and their proficiency level. This paper firstly reviews the study on vowel discrimination and the one on the quality of vowel reduction, and then introduces another two studies: the one is on pronunciation error, and the other is on pronunciation modeling of reduced vowel.

1 Speech data and rating procedure

Each participant out of 101 Asian English learners was recorded as they read a passage aloud. The passage was a fable of Aesop, "The North Wind and the Sun" (see Appendix A). The group was composed of forty Japanese, seventeen Chinese, nineteen Korean, six Filipino, ten Thai, four Vietnamese, four Cambodians, and one Indonesian. These participants were either undergraduate or graduate students. Five raters joined this evaluation; they were Japanese language teachers who had participated in the rater training based on Common European Framework of References (CEFR), and their reliability had been examined by Generalilzability Theory and Multi-faceted Rasch Analysis (MFRA). The raters evaluated all the speeches that were read by the 101 Asian English learners by using fourteen evaluation items with 6 point scale (see Appendix B), and in the studies reported here, read-aloud speeches were randomly selected and used to investigate the relationship between the evaluation scores and pronunciation.

2 Vowel discrimination

Kitagawa, Kondo, and Nakano (2007) explored the relationship between vowel discrimination rate and evaluation scores, focusing on the difference in vowel systems between the target language and the first language (Japanese).

Speech spoken by thirty-eight Japanese learners of English was analyzed. They were divided into three groups: high, mid, and low levels, according to their evaluation score. Data from five participants were randomly selected from each group and acoustically measured. This decision was made because only five Japanese learners belonged to the high level.

Taking account of the differences and the similarities between the Japanese and English vowel systems, three pairs of vowels were chosen: /ɪ/ and /i/, /u/ and /ʊ/, and /æ/ and /ʌ/. The Japanese have five vowels in quality, and each vowel can be both long and short. On the other hand, English has a much larger vowel system with about eleven or twelve vowels. Assuming the interlanguage transfer of vowels, the Japanese tend to produce two or more distinct English vowels with one Japanese vowel. For this reason, the three targeted pairs of

English vowels are considered to be less distinguishable for Japanese English learners.

Acoustic measurements were performed with the acoustic analysis software, Praat. First, segmentation was provided to each speech. Then, the F1 and the F2 of the target vowel were measured at the point that was considered to be under the least influence from the sounds adjacent to it. This point was visually defined by hand with the help of formant tracks from F1 to F5. The words including the target vowel were as follows (The number in the brackets indicates how many times the word was spoken): agreed [1], succeed [1], and immediately [1] for /i/; wind [4], which [1], and considered [1] for /ɪ/; disputing [1], blew [2], and two [1] for /u/; should [1], could [1], and took [1] for /ʊ/; traveler [4], wrapped [1], and last [1] for /æ/; and sun [3], one [1], other [1], and up [1] for /ʌ/. If a word was repeated, the F1 and F2 values were averaged; as a result, the data of each target vowel was obtained from three or four different words respectively. Whether or not each speaker differentiated each vowel from another was examined as a physical reference of the achievement in vowel quality. A statistical test, a discriminant analysis, was conducted to investigate the speakers' achievements in classifying these six vowels.

Tables 1, 2, and 3 show the results of the linear discriminant analysis with cross validation between every possible pair for each group according to overall pronunciation proficiency.

Table 1: The ratio of correct classifications (High)

	/ɪ/	/u/	/ʊ/	/æ/	/ʌ/
/i/	86.7	90.0	90.0	100	100
/ɪ/	-	70.0	80.0	100	100
/u/	-	-	63.3	96.7	100
/ʊ/	-	-	-	96.7	100
/æ/	-	-	-	-	62.9

Table 2: The ratio of correct classifications (Mid)

	/ɪ/	/u/	/ʊ/	/æ/	/ʌ/
/i/	51.4	80.0	83.6	100	100
/ɪ/	-	75.9	75.9	96.4	97.1
/u/	-	-	63.3	96.6	94.3
/ʊ/	-	-	-	96.6	91.4
/æ/	-	-	-	-	53.7

Table 3: The ratio of correct classifications (Low)

	/ɪ/	/u/	/ʊ/	/æ/	/ʌ/
/i/	43.3	90.0	76.7	100	97.1
/ɪ/	-	90.0	83.3	100	100
/u/	-	-	56.7	93.1	91.4
/ʊ/	-	-	-	93.1	91.4
/æ/	-	-	-	-	55.9

The correct classification rates between the target pairs (i.e., /i/ and /ɪ/, /ʊ/ and /u/, and /æ/ and /ʌ/, whose counterparts in Japanese are /i/, /u/, and /a/, respectively) were fairly low for all the groups; however, the speakers in the high-level group tended to succeed in classifying /i/ and /ɪ/ (86.7%).

Although these results indicate the Japanese learners' vowel pronunciation features, they cannot be the factor that distinguishes the learners' read-aloud speech levels. Although the examinees' discrimination rate of /i/ and /ɪ/ at the high-levels is very high, especially when compared with the examinees in the other two levels, the discrimination rates are not ideally suited for the other vowel pairs. Ideally, the data would have shown that the high-level examinees obtained higher discrimination rates, the mid-level obtained moderate rates, and the low-level obtained the lowest rates.

3 Vowel reduction

English reduced vowels in unstressed syllables belong to one of the three vowels: /ɪ/, /ʊ/, and /ə/. According to Roach (2000), these unstressed vowels, or weak syllables, are likely to be shorter in duration and have lower intensity and different qualities than stressed (strong) syllables. On the contrary, in the Japanese prosodic system, a change of F0 is required in order to realize the accent and the long-short contrast in sound length to achieve the mora duration. Furthermore, variations in intensity and vowel quality are not necessary in Japanese phonology. In consideration of these differences between the English and Japanese phonological systems, it is hypothesized that these features of reduced English vowels can be adopted as the indices that categorize a learner's proficiency when reading aloud.

Kitagawa and Kondo (2008) only focused on /ə/ as the target vowel, and this vowel was analyzed with the acoustic analysis software, Praat. The test-tokens were /ə/ in the words "attempt," "around," "agreed," "along," "considered," "confess," and "obliged." The underlined vowels are supposed to be produced as /ə/. First, each participant's speech was segmented. Then, the target features, F0, duration, intensity, and F1 and F2 were measured. With regard to the first three properties, the values of stressed vowels within the same word (the bolded vowels) were also analyzed in order to calculate their relative values, and the ratios of unstressed vowels to stressed vowels were obtained. The F1 and the F2 of the target vowels were measured at the point that was considered to be under the least influence from adjacent sounds. This point was visually defined by hand with the help of formant tracks from F1 to F5. Additionally, these values were normalized in order to compare

the data across the speakers based on Guion's method (Guion, 2003). In Guion's normalization, first, one speaker's F3 value for /æ/ is taken as a norm, because F3 is commonly recognized as a reflection of vocal tract length. Second, this norm F3 value is divided by the mean F3 values for /æ/ produced by each speaker, and the factor for each speaker is calculated.

The speech data were grouped into three levels, high, mid, and low, and five data sets were randomly selected from each group and were acoustically measured; each group was comprised of two males and three females.

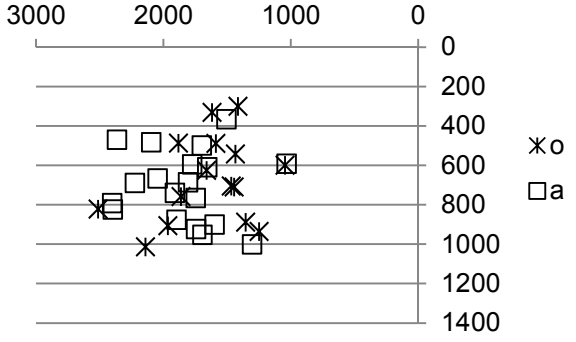

Figure 1: F1 and F2 of the Reduced Vowels at High-level

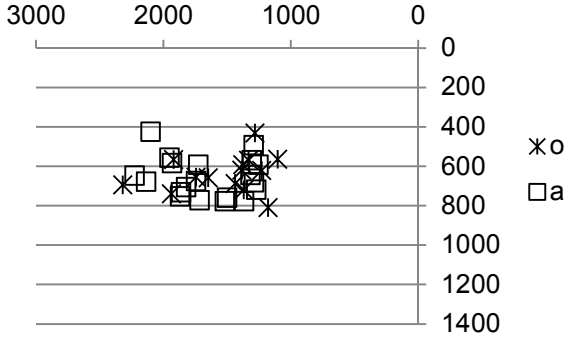

Figure 2: F1 and F2 of the Reduced Vowels at Mid-level

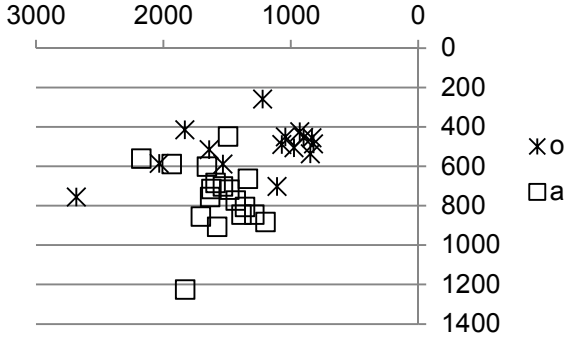

Figure 3: F1 and F2 of the Reduced Vowels at Low-level

Finally, the F1 and F2 values are multiplied by this respective factor. In this study for instance, speaker A's average F3 value for /æ/, 2696.42, was taken as a norm. Given speaker B's average F3 value for /æ/, 2275.35, the factor for speaker B was 1.185 (2696.42 divided by 2275.35). Then the normalized F1 and F2 values of speaker B were obtained by multiplying each with 1.185. These formant values were transformed to mel scale in order to examine the perceptual vowel quality. Mel scale is a perceptual scale of sound pitch. The difference in mel scale indicates the difference of which a human being senses sound pitch. This scale is defined as:

$$\mathrm{Mel(f)}2595\log_{10}(1+\frac{\mathrm{f}}{700}) \qquad (6.1)$$

where f is hertz (Young, Evermann, Gales, Hain, Kershaw, Liu, Moore, Odell, Ollason, Povey, Valtchev, and Woodland., 2006).

In Figures 1, 2, and 3, F1 and F2 values are plotted depending on the orthographic spelling. As far as the visual observation goes, it can be noted that unstressed vowels were more centralized in their quality for the high-level and mid-level groups than the low-level group. Reduced vowels produced by the low-level group were more separated in the vowel space according to the spelling, divided into "a" space and "o" space. The fact that the mid-level group probably performed better than the high-level group may propose that the accuracy of vowel quality reduction does not matter after a certain degree of accuracy is satisfied.

Table 4: Mean and standard deviation of intensity

	High	Mid	Low
Mean	-2.02	-2.09	-1.51
S.D.	3.99	3.41	3.57

Table 5: Mean and standard deviation of F0

	High	Mid	Low
Mean	.92	.91	.92
S.D.	.07	.08	.10

Table 6: Mean and standard deviation of duration

	High	Mid	Low
Mean	.47	.52	.57
S.D.	.29	.25	.27

The means and standard deviations of intensity, F0, and duration are shown in Tables 4, 5, and 6. For intensity, the log ratios calculated by subtracting the intensity of stressed vowels from that of unstressed vowels are shown in Table 4. If the value was less than 0, the intensity of the unstressed vowels was successfully lower. For F0 and duration, the ratio of unstressed syllables to stressed syllables is shown in Table 5. If the value is less than 1, the F0 of the unstressed vowel is successfully lower and the

duration is successfully shorter.

As shown in Table 4, the intensity of the unstressed vowels produced by the learners in all groups was weaker than stressed vowels, but there was no significant difference between groups. This was also the case with the analysis of F0, where little difference was found among the groups. The results indicate that these features cannot be used as a predictor of the evaluation score when compared with the duration shown in Table 6.

4 Pronunciation error

Kondo (2010) investigated the relationship between the overall evaluation and the number of pronunciation error. Thirty-one speech data were randomly selected and were analyzed in terms of pronunciation error.

As the index of pronunciation error, the sum of the number of elisions, epentheses, and replacements were examined. An example of an elision found in this analysis is the deletion of consonants (e.g., in the word "succeeded" /səksidɪd/ → /səsidɪd/). There was no deletion of vowels found in the present data. Examples of epenthesis were observed at the end of words that end with plosive sounds such as "first," "wind," and "fold." Replacement was defined as the replacement of vowels and consonants. Examples of the pronunciation errors found in this analysis were the replacement of words (e.g., cloak → coat), the replacement of vowels (e.g., in the word "obliged" /aɪ/ → /ɪ/), and the replacement of consonants (e.g., in the word "first" /f/ → /p/). Through the observation of the data, no addition of consonants was found. The measurements of these features were done with the acoustic analysis software, Wavesurfer (Sjölander and Beskow, 2000). Regarding the measurement of pronunciation errors, each point was visually defined by hand with the help of formant tracks from F1 to F5.

The average number of pronunciation error was 2.6 with 3.0 S.D. A negative correlation was found between the number of pronunciation error and the evaluation score in the present data (-.68).

5 Acoustic likelihood computation

Kondo and Nakano (2009) constructed an automatic scoring system for L2 read-aloud speech. The system measures examinees' speech characteristics, which are found to be statistically significant predictors of the evaluation score, the indices of speech rate and of rhythm and predicts the scores. To measure examinees' speech characteristics, the system adopts the Hidden Markov Model Toolkit (HTK), which is a tool for Hidden Markov Model (HMM) that has been optimized for speech recognition (Young, Evermann, Gales, Hain, Kershaw, Liu, Moore, Odell, Ollarson, Povey, Valtchev, and Woodland, 2006). To train an acoustic model for automatic speech recognition (ASR), the speech data from TIMIT Acoustic-Phonetic Continuous Speech Corpus (Garofolo, Lamel, Fisher, Fiscus, Pallett, Dahlgren, and Zue: 1993) were used to train the HMM. To adapt the model to English spoken by Asian learners, the speech data of the read-aloud speech of 101 Asian English learners were used. In the process of adopting and training the model, HTK phone-aligned the target speech data based on the order of occurrence of phones by referring to the text labels and the pronunciation dictionary. HTK must run through model training several times to create robust HMMs. A gender-independent HMM recognizer was bootstrapped to native speech data and was trained by using non-native speech data. Phonetic time alignments were generated for the speech data using the Viterbi algorithm with the native English model (TIMIT Acoustic-Phonetic Continuous Speech Corpus) trained with the speech data of Asian learners of English. Through this process, phone-aligned speech data are obtained.

Using HTK for ASR, we obtained the beginning and ending times when a phone is uttered and its acoustic likelihood score, which is an index of distance between the phone uttered and the phone in the model. In the present study, adopting the acoustic log-likelihood score, the quality of reduced vowels was examined.

Twenty examinees took the test on this system, and three raters and the system evaluate the examinees read-aloud speech. The raters give categorical score (A, B, and C) to the examinees. The system predicts examinees' score by using two indices: speech rate and rhythm, and gives categorical score: A, B, and C. In the analysis, the score A is converted to 3; B, to 2; and C, to 1, and the sum of the scores given by three human raters and the system was used as each examinee's score.

The target phones in the present study are /ə/ in the words "attempt," "around," "agreed," "along," "considered," "confess," and "obliged" whose underlined vowels are supposed to be produced as /ə/. The average of acoustic likelihood score of these seven target phones was calculated for each examinee.

A moderate correlation coefficient (.61) was found between the sum of the score and the average of the acoustic likelihood score.

6 Discussion and conclusion

Through the studies reported here, we have made attempts to investigate to what extent learners' pronunciation has influence on overall evaluation of

read-aloud speech in an Asian context.

In Kitagawa, Kondo, and Nakano (2007), we grouped the learners into three levels: High, mid, and low, and examined the vowel discrimination rates, mainly focusing on three vowel pairs: /i/ vs. /ɪ/, and /u/ vs. /ʊ/, and /ʌ/ vs. /æ/, but the rates were found to be poor predictors of the learners' levels.

Kitagawa and Kondo (2008) revealed that the characteristics of reduced vowel did not distinguish among the learners' level.

Kondo (2010) examined the relationship between the number of pronunciation error defined in this study and the evaluation score, found the moderate correlation (-.68). This implies that less number of pronunciation error an examinee makes, higher score he/she obtains. However, because the correlation coefficient is not high, the influence of the number of pronunciation error is limited on the overall evaluation of read-aloud speech.

Finally, we analyzed the speech data from Kondo and Nakano (2009), and examined the relationship between the overall evaluations and acoustic log-likelihood scores. The log-likelihood scores were found to be moderately correlated with the overall evaluation scores (.61).

The results of the researches reported here indicate that learner's vowel quality has some influence on overall evaluation in read-aloud speech in the Asian context. However, vowel quality does not have a great amount of influence on overall evaluation in read-aloud speech.

Acknowledgment

This work was supported by Grant-in-Aid for Scientific Research (C) (23520721).

Appendix A: The North Wind and the Sun

The North Wind and the Sun were disputing which was the stronger when a traveler came along wrapped in a warm cloak. They agreed that the one who first succeeded in making the traveler take his cloak off should be considered stronger than the other.

Then the North Wind blew as hard as he could, but the more he blew the more closely did the traveler fold his cloak around him; and at last the North Wind gave up the attempt.

Then the Sun shone out warmly, and immediately the traveler took off his cloak. And so the North Wind was obliged to confess that the Sun was the stronger of the two.

Appendix B: Evaluation items

1. Loudness
2. Sound pitch
3. Quality of vowels
4. Quality of consonants
5. Epenthesis
6. Elision
7. Word stress
8. Sentence stress
9. Rhythm
10. Intonation
11. Speech rate
12. Fluency
13. Place of pauses
14. Frequency of pauses

References

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. D., Dahlgren, N. L., & Zue, V. (1993). TIMIT Acoustic-Phonetic Continuous Speech Corpus. Philadelphia: Linguistic Data Consortium.

Guion, S.G. (2003). The vowel systems of Quichua-Spanish bilinguals: An investigation into age of acquisition effects on the mutual influence of the first and second languages. *Phonetica* 60, 98-128.

Kitagawa, A., Kondo, Y., & Nakano, M. (2007). Does vowel quality matter? *Proceedings of 12th Conference of Pan-Pacific Association of Applied Linguistics*, 224-227.

Kitagawa, A., & Kondo, Y. (2008). Reduction of vowels by Japanese learners of English. *Proceedings of 13th Conference of Pan-Pacific Association of Applied Linguistics*, 227-230.

Kondo, Y., & Nakano, M. (2009). Construction and implementation of automatic L2 speech evaluation system. *Proceedings of 14th Conference of Pan-pacific Association of Applied Linguistics*, 33-38.

Kondo, Y. (2010). *The development of automatic speech evaluation system for learners of English*. Unpublished doctoral dissertation, Waseda University, Tokyo, Japan.

Kondo, Y., Tsutsui, E., Nakano, M., Tsubaki, H., Nakamura, S., & Sagisaka, M. (2007). "The relationship between subjective evaluation and objective measurements in Second language oral reading" [Eigo gakushusha ni yoru ondoku ni okeru shukanteki hyoka to kyakkanteki sokuteichi no kankei]. *Proceedings of the 21st General Meeting of the Phonetic Society of Japan*. 51-55.

Kondo, Y., Tsutsui, E., Tsubaki, H., Nakamura, S., Sagisaka, Y., & Nakano, M. (2007). Examining predictors of second language speech evaluation. *Proceedings of 12th Conference of Pan-Pacific Association of Applied Linguistics*, 176-179.

Nakano, M., Kondo, N, & Tsutsui, E. (2008). Fundamental Research on Automatic

Speech Evaluation. *9th APRU Distance Learning and the Internet Conference--New Directions for Inter-institutional Collaboration: Assessment & Evaluation in Cyber Learning*. 207-212.

Kondo, Y., Tsutsui, E., & Nakano, M. (2010). Bridging the Gap between L2 Research and Classroom Practice (2): Evaluation of Automatic Scoring System for L2 Speech. *Proceedings of INTERSPEECH 2010 Satellite Workshop on Second Language Studies*. CD-ROM

Roach, P. (2000). *English phonetics and phonology: A practical course*. Cambridge: CUP.

Sjölander, K., & Beskow, J. (2000). WaveSurfer - an open source speech tool. In Yuan, B., Huang, T., & Tang, X. (Eds.), *Proceedings of ICSLP 2000, 6th Intl Conf on Spoken Language Processing*, 464-467. Beijing.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. (2006). *HTK book*. Cambridge University Engineering Department.