The Evaluation of L2 Pronunciation by the Raters' Language Background*

Seokhan Kang
(Seoul National University)

Kang, Seokhan. (2013). The Evaluation of L2 Pronunciation by the Raters' Language Background. *The Linguistic Association of Korea Journal, 21*(4), 99-117.
Evaluating L2 fluent pronunciation causes some problems for L2 teachers because of its subjectivity in evaluation. This paper is trying to suggest an educational implication for nonnative English pronunciation teaching. A speaking test is designed and administered to Korean undergraduates to determine which measured cues raters of natives and nonnatives associate with their rating patterns of L2 English proficiency. Fourteen raters (7 natives, 7 nonnative) are recruited to rate 30 participants. The results clearly show that the measured parameters such as pitch range, speech rate, and pause duration cues produced by Korean English learners are evaluated differently from raters' language background. This paper argues for the need to import the parameter values fitted for the objective evaluation of L2 English pronunciation.

Key Words: speaking evaluation, fluency, second language acquisition, English pronunciation test, phonetic cues

1. Introduction

Evaluating English oral proficiency could be a critical problem for L2 English educators because of its subjectivity in evaluation (Saville and Hargreves, 1999; Kang and Ahn, 2010). Saville and Hargreves (1999) report that

* This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2012S1A5B5A02025277).

raters are hard to keep clear-cut criteria consistently for every testee and also sometimes the criteria could vary depending on testees or situations. Kang and Ahn (2010) suggest that this subjective evaluation raises the critical problems for L2 speech by L2 raters.

Especially the issue of valid and/or reliable evaluation of L2 learners' English oral proficiency by nonnative English raters could be heated under the curriculum emphasizing on communicative competence-oriented English learning and teaching (e.g., Luoma, 2004; Kang, 2013). Thus, it is meaningful to search for an objective way in evaluating L2 English oral proficiency by nonnative English raters. With this particular aim, this paper is trying to seek for why two rating groups (native and non-native rating groups) differ in evaluating L2 pronunciation and suggest a way of objective measurement of fluency parameters. This suggestion is based on the results of rating experiments, comparing both groups of native English speakers and nonnative English raters of Korean.

Oral fluency is one aspect of communicative success, along with comprehensibility (Derwing, Thompson, Monro, 2006). In L2 acquisition area, fluent pronunciation refers to phonetically various features such as frequency, duration, intensity, stress, and prominence in both segmental and surpasegmental sides. Thus, it is well-established that fluent L2 pronunciation is essential in L2 acquisition (Segalowitz, 2010). According to some previous studies in second language acquisition (Aoyama, Guion, Flege, Tsuneo, & Akahane-Yamada, 2008; Guion, Flege, & Loftin, 2000; Johns-Lewis, 1986), fluent pronunciation is determined by some prominent features in L2 phonetics such as pitch range, speech rate, boundary cues, pause frequency and duration, or declination tilt. In this study, the phonetic cues of pitch range, speech rate, and pause duration are analyzed because these threes are most prominent factors to decide on fluency or intelligibility of L2 speaking (Alberchsten, Henriksen, & Faerch, 1980; Munro & Derwing, 1996; Tromfimovich & Baker, 2006). This study is trying to suggest a valid/objective evaluation method for L2 pronunciation.

1.1 L1 Influence

The evaluation of L2 speech is greatly influenced by raters' L1 language

background (Munro & Derwing, 1996). In this vein, it is important to understand the prosodic as well as the segmental structures of Korean and English in this study. Korean has different prosody features from English, in which it has two prosodic units above the prosodic word: the intonational phrase (IP) and the accentual phrase (AP) (Jun, 2005). An IP is defined by phrase final lengthening as the form of a boundary tone and also is the highest prosodic unit defined by intonation. APs in Korean have some predictable pitch accents related with stressed syllables in their domain and also lack the phrase tone which occurs at the end of the accentual phrase. In contrast, English is a stress language in which one syllable is stressed within the prosodic unit. The stressed syllable tends to produce a greater duration, higher pitch, and more complicated pitch contour than the unstressed syllables. English has two prosodic units above the prosodic foot: the intonation phrase (IP) and the intermediate phrase (iP). An IP is the highest prosodic unit defined by intonation and may contain one or more iPs. It has final lengthening with the final falling F0 in the case of statement sentences.

The structural difference of L1 prosody clearly influences on L2 fluent speech. It is well known that native-like L2 speech is defined as the listeners' judgment of how natural an utterance sounds, spoken without undue pauses, filled pauses, hesitations, slow speech rate, or dysfluencies (Derwing and Munro, 1997). Derwing and Munro (1996) argue that L2 learners' mother tongue influences judging the native-likeness of the target language. Based on the previous research, to what extend L1 factors influence on the L2 speech evaluation becomes main issue in our L2 pronunciation assessment.

1.2. The purpose of the study

The current study investigates L2 speech evaluation by the raters' background language. The goal of the study is to extend our understanding of factors influencing the L2 fluency evaluation by both L1 rating groups – native English raters, and native Korean raters of L2 English. Experiment checks the hypothesis that the evaluation of good or bad L2 pronunciation is different by the effect of raters' language background. Acoustic analysis of L2 suprasegmental production is carried out focusing on to what extent the features

of the acoustic cues change over both groups of native and nonnative raters. Three fluency cues (F0 range, speech rate, and syllable duration) which proved to be the most important cues to determine the L2 fluency ratings (e.g., Munro and Derwing, 1996; Mennen, 2006; Trofimovich and Baker, 2006), are investigated.

(1) Do non-native Korean raters on nonnative English speaking proficiency by Korean exert distinctive evaluating patterns on good or bad L2 pronunciation?
(2) How do both rating groups (native and nonnative English rating groups) differ in drawing on evaluation distribution on nonnative English speaking?

The primary research goal aims to investigate commonalities and the degree to which the goodness decision differs for native and non-native raters, and by implication, the justification for both groups beyond acoustic analysis. Thus, preliminary findings are presented on L1 correlates of L2 proficiency evaluation to mitigate the potential problems on validity/objectivity by non-native raters.

2. Experiment

2.1. Participants

Test-takers of thirty Korean male students were selected from a college-level English class of a university in Seoul, Korea, and were informed about the research project. The students were assigned to mid-level classes based on the English proficiency test before beginning the regular classes of the semester. The college sorted students into one of three class levels according to the scores of a placement test measuring four English skills (reading, writing, speaking, listening). Thus, each test-taker in this study would be considered as almost equal level of English proficiency.

The experiment was administered in a computer-mediated interview format for the purpose of study. The semi-directed method was chosen because of its

effectiveness, reliability, and easy accessibility (e.g., Kim, 2006; Shin, 2002; Kang & Ahn, 2010). The test-takers' voices were recorded in a quiet room and it lasted approximately 30 minutes. After the assessment was conducted after two weeks of speaking tests, both groups of native and non-native raters scored the speaker' speech individually.

Fourteen raters (seven native English raters and seven non-native Korean raters) participated in the study. In order to confirm that raters were sufficiently qualified, certain participation criteria were applied: (1) at least two years of teaching experience in English for nonnative Korean students; and (2) at least master degree in a field related with linguistics or English education. The two groups were similar in that they were all university English teachers. Most of the Korean raters, 3 males and 4 females, hold master or doctoral degree in the field of English linguistics or education, and have experience in teaching English in the universities for 4 to 12 years, ranged from 34 to 45 years old. They have no experience in learning English in English-speaking countries over 6 months. Also most of the native English raters, 4 males and 3 females, hold master or doctoral degree, and have experience in teaching English for 5 to 7 years, ranged from 31 to 38 years old. Small compensatory money was paid.

2.2. Materials and Procedures

The speech scripts were presented to Korean subjects over a loudspeaker using a laptop computer in question-answer-question sequences. A delay of around ten seconds was provided after the second question in each sequence, allowing some time for the production of the target sentences. It was intended to avoid the direct imitation from the sensory memory (e.g., Flege and Fletcher, 1992; Flege, 2006). The sequences were presented randomly. The elicitation of the second repetition was analyzed. Before they produced the sentences, it was confirmed that they knew what the sentences meant, and that they knew how to pronounce them. Following are the sentences produced by subjects:

1. A: What did Mary do?
   B: Mary met him at the same place.
2. A: What does you do?

B: Raise your right hand, if the pastor calls your name.
3. A: What did he do?
B: With a light hammer, the carpenter hit the nail.
4. A: What did he do?
B: All of a sudden, the policeman rushed to the market.
5. A: What did they do?
B: People couldn't sleep well last night, because of the noise.

Each participant was asked to read each English sentence one time. Before they produced the sentences, it was confirmed that they did know what the sentences meant, and they knew how to pronounce them. Also Korean subjects were given 30 minutes to practice the sentences before the experiment. The sounds were recorded by a SONY TASCAM DA P-1 DAT recorder with Schure SM 10A microphone, and digitalized in 44.05 kHz and 16 bit resolution.

2.3. Measured cues

Among the various phonetic features, our focus has been on parameters of the F0 range, speech rate, and pause duration because these three factors are closely related with fluency-based suprasegmentals (Aoyama et al., 2008; Guion, Flege, Liu, Yeni-Komshian, & Grace, 2000; Trofimovich & Baker, 2007). Generally these cues affect listeners' ratings of foreign accent in L2 speech and are viewed as determinants of both fluency and intelligibility (Alberchsten, Henriksen, & Faerch, 1980; Munro & Derwing, 1996). The cues are measured as follows:

F0 range: F0 range was known to be a good signal to measure English proficiency (Backman 1979; Willems 1982). Generally the beginners of English exhibited a narrower F0 range. In this study, the range was measured from the highest point to the lowest point of the fundamental frequency. This study used the F0 track, and also the wave forms associated with the vibration of the vocal folds as a supplementary check.

Speech rate: The speech rate proved to be a good evaluator which would

help decide native-like pronunciation (Derwing & Munro, 1996; Guion, Flege, Liu, Yeni-Komshian, & Grace. 2000). In this study, speech rate was measured from the initial acoustic signal in both waveform and spectrogram to the final acoustic or spectral cues of boundary.

Pause duration: The duration of pause was associated with fluency-based suprasegmentals (Tromfimovich & Baker, 2007). Pause duration affected listeners' ratings of foreign accent in L2 speech and were viewed as determinants of both fluency intelligibility (Albrechsten, Henriksen, & Faerch, 1980). In this study, pause duration was measured between phrases in periodic sentences.

2.4 Research design and Data analysis

This study applied mixed analysis methods which could provide a depth and breadth in that a single approach might lack by itself. In particular, mixed analysis design was set up to understand a research problem more completely. The design was conducted in two steps: an initial phase of Rasch method and then followed by a phase of ANOVA analysis. An advantage of the design was that a researcher could explain more clearly on how the Rasch findings helped elaborate or extend the ANOVA results. A separate section in this study might discuss how the two phases were connected in the research process.

The data consisted of 420 valid ratings, awarded by 14 raters to fluency task responses by 30 test-takers. Each rater rated every student's performance on every task, so that the data matrix was fully crossed. The Rasch method was conducted for why some criteria showed wider difference between native and nonnative groups and then analyzed by RM ANOVAs as follows:

a. Rasch analysis

The data were analyzed using a multifaceted Rasch measurement (MFRM) approach. This Rasch method allowed for including many aspects, or facets in the ratings (Bachman, 2004; Winke et al., 2012). To conduct a MFRM analysis, the FACETS program (Version 3.64; Lincare, 2008) was used. It used the scores in that raters were aware of examinees based on each of the five criteria (i.e.,

grammar usage, vocabulary usage, cohesive discourse, clear pronunciation, fluency) to analyze raters' severity depending on L1 group, raters' consistency, task measurement difficulties, and test-takers severities. The rater facet was entered as a dummy facet and anchored at zero. A Many-faceted Rasch Measurement Model was used to analyze the data. The formular could be built up like follows;

$$\log \frac{pnijlpk^{k-1}}{pnijlpk^{k}}$$ = English speaking fluency

- fluency of examinee n
- difficulty of criterion i
- difficulty of task measurement l
- severity of rater j
- rater's severity by L1 p
- difficulty of receiving a rating of k relative to a rating of k-1

b. Statistical analysis

These measures were analyzed with Repeated Measures of Analyses of Variance (RM ANOVAs) which were conducted for statistical evaluation of the groups with the following parameters: Dependent variables of fundamental frequency, speech rate, and pause duration were examined by the factor of Group (native and nonnative English raters). The repeated measure of L2 speeches was used in order to consider the individual variation (each sentence by each speaker) along with within group variation (F0 range, speech rate, and pause duration). Repeated measures were used in order to account for within speaker variance in pronunciation. Its design is able to factor out some of the variation that occurs within individuals.

3. Results

3.1. Reliability analysis

Consistency of the ratings for each group was estimated by Cronbach's

coefficient alpha, which reflected the level of agreement within each group as a whole. The results showed that the ratings of native raters showed a slightly greater reliability (coefficient α = .81) than the nonnative rates (coefficient α = .71) indicating that the nonnative raters share the different consistency of the ratings as native speakers. To check the inter-rater reliability for both groups, Pearson correlation coefficients were estimated to examine the agreement among the raters in each group. In the case of the nonnative raters, the average Pearson product moment correlation coefficient for fluency is .73., while the average correlation coefficients among the native raters were slightly higher than those given by the nonnative raters of fluency as .86.

3.2. Raters severity and consistency

Table 1 shows a summary of the rater measurement report from the FACETS analysis indicating the degree of severity graded by each rater. Raters were ranked from most severe to most lenient by groups; the higher the rater severity measure, the more severe the rater. In this study, similarity between both rating groups could be found in severity measurement as shown in 1.88 of mean logit for English raters and 1.77 of mean logit for Korean raters. Thus, the result supported the non-native raters' credibility, meaning that two different groups of native and non-native raters exhibited similar strict scores in overall ratings. Following were measurement results for the rater facet:

Table 1. Measurement results for the rater facet

Group	Raters	Obs. average	logits	measure S.E.	Infit MnSq	Outfit MnSq	PtBis
English raters	1	1.84	1.73	0.09	1.08	1.06	0.51
	2	1.95	1.82	0.09	0.76	0.76	0.48
	3	2.28	2.21	0.09	0.77	0.77	0.50
	4	1.54	1.64	0.09	0.64	0.63	0.57
	5	1.96	1.86	0.09	0.82	0.82	0.52
	6	1.83	1.72	0.09	0.92	0.71	0.49
	7	1.92	1.83	0.09	0.77	0.83	0.50
	Mean	1.88	1.88	0.09	1.07	1.06	0.56
Korean raters	1	1.58	1.68	0.11	0.64	0.66	0.46
	2	2.03	2.14	0.12	1.44	1.44	0.48

	3	1.73	1.83	0.11	0.69	0.68	0.47
	4	1.62	1.88	0.09	0.73	0.74	0.47
	5	1.64	1.69	0.09	1.35	1.35	0.59
	6	1.74	1.82	0.12	0.71	0.72	0.51
	7	1.82	1.92	0.12	1.09	0.89	0.48
	Mean	1.77	1.79	0.10	0.93	0.92	0.52

The result suggests that two rating groups showed a little difference in assessing oral proficiency skills. It means that non-nativeness was a significant variable to determine the strictness of the rating patterns. To examine rater's consistency, the infit indices of each rater were assessed. Raters' fit statistics indicated the degree to which each rater was internally consistent in his ratings. Although a proper range of infit mean squares for raters was flexible depending on researchers (Myford and Wolfe, 2004; Wright and Linacre, 1994), this study was set at 0.5 and 1.5 respectively as the lower and upper quality control by following studies of Kim (2009) and Zhang and Elder (2011). In case of raters' consistency test, infit mean squares of raters were favored rather than outfit squares of raters because of its consistency (Choi, 2011). Infit mean square values greater than 1.5 indicated significant misfit, or a high degree of inconsistency in the ratings, while infit mean square values less than 0.5 indicated overfit, or a lack of variability in their scoring. The fit statistics in the study reported that that none of raters, regardless of L1 language backgrounds, showed misfit or overfit ratings, meaning that all raters kept consistent rating patterns. In short summary, non-native English raters were consistent in their ratings, although they exhibited somewhat lenient ratings. Overall, it was safe to say that non-native raters' grading showed a little different patterns with native English raters in evaluating L2 fluency.

4. Acoustic analysis

The sentences were analyzed to investigate differences of prosody goodness between the two groups. Table 2 presents mean values and standard deviations (in parenthesis) of F0 range, speech rate, and pause duration for good speaking (30%) evaluated by both groups.

Table 2. The measured cues of the fluent speech by groups.

	Naive E.	Non-native K.
F0 range (Hz)	120 (55)	94 (43)
Speech rate	1.71 (0.3)	2.34 (0.45)
pause duration	0.19 (0.11)	0.39 (0.13)

4.1. F0 Range

The RM analysis of variance confirmed that there was a significant effect of group on overall F0 range, $F(1, 125) = 19.278$, $p < .001$. Tukey's tests ($p < .05$) revealed that the F0 range was smaller for the nonnative raters than the naive groups. Figure 1 presents the F0 range evaluated by both groups. The result showed that Korean raters graded good speech based on a comparatively smaller F0 range. Korean raters' evaluation based on a narrower F0 range could be evidence of the influence of the native-language (e.g., Scherer, 2000) as well as a foreign accent (e.g., Backman, 1979; Willems, 1982). Note that the F0 range in the Korean language exhibited 110 to 200 Hz in males (Lee, 2003), while native English speakers showed 60 to 240 Hz in males (Crutenden, 1997).

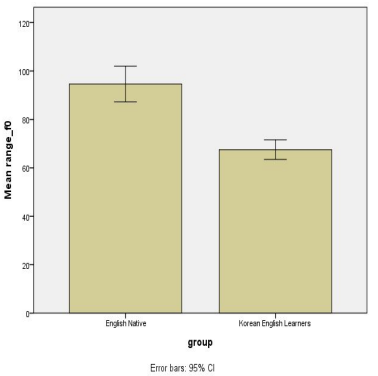
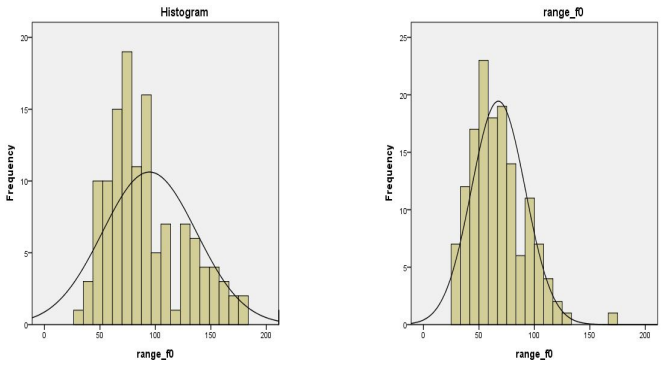


Figure 1. F0 range of good pronunciation evaluated by natives (left) and nonnatives (right)

For more depth analysis, the relationship between F0 range and good speech evaluation is investigated as follows:



(a) native English raters (b) Korean L2 raters

Figure 2. Two histograms of F0 range evaluated by both groups.

Good speech of 30% evaluated by native English speakers is distributed within 70 Hz and 60% from 71 Hz to 91 Hz of the range. On the contrary, good speech of 30% evaluated by Korean English learners exists within 52 Hz and 60% from 53 Hz to 72 Hz of the range. Considering that median values of the F0 range are 65 Hz for Korean English learners and 85 Hz for native English speakers, the frequent distribution of data between two groups is clearly different. It indicated that F0 range could be the solid indicator to the objective assessment.

4.2. Speech Rate

The results of the RM analysis of variance confirmed a significant effect of group on speech rate, $F(1, 125) = 2.234$, $p < .05$. Tukey's tests ($p < .05$) revealed that the speech rater was shorter for the native group, and longer for nonnative group. Figure 3 presents the mean value of duration for the three groups.

Out of the various parameters for prosody, durational speech rate might be regarded as the most salient and reliable feature (Adams & Munro, 1978; Sluijter & van Heuven, 1996). Obviously, the speech rate could be understood as durational length because as speech rate increased, the durational length naturally decreased. Generally, the native-like English learners spoke faster than English beginners (Guion, Flege, Liu, Yeni-Komshian, & Grace, 2000; Munro &

Derwing, 1996) and a slower speech rate was possibly related to foreigner's pronunciation (Derwing & Munro, 1997).

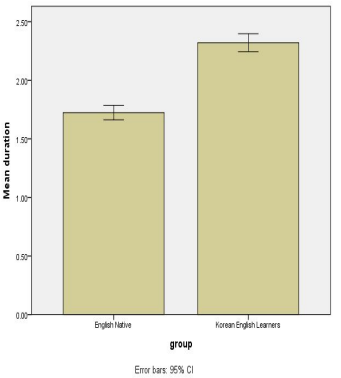
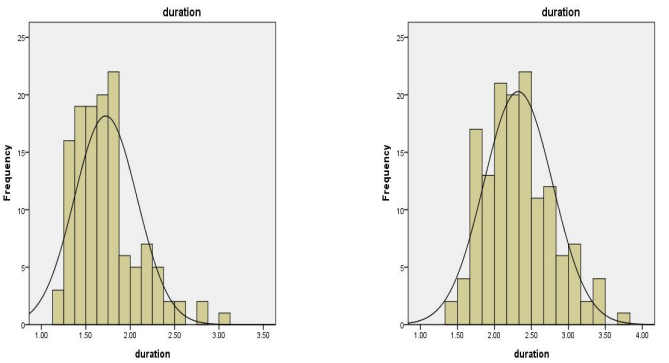


Figure 3. F0 range of good pronunciation evaluated by natives (left) and nonnatives (right)

A slower speech rate evaluated by Korean raters could be evidence of the influence of the native-language (Aoyama & Guion, 2008). Thus, speech rate is an important factor when nonnative raters decided fluent pronunciation. For more depth analysis, the relationship between speech rate and good speech evaluation is investigated as follows:



(a) native English raters (b) Korean L2 raters

Figure 4. Two histograms of speech rate evaluated by both groups.

Good speech of 30% evaluated by native English speakers show 1.49 seconds

and 60% show 1.78 seconds. On the contrary, good speech of 30% by Korean raters hold 2.08 seconds and 60% keep 2.34 seconds. Considering that median values of the speech rate are 2.28 seconds for Korean English raters and 1.65 seconds for native English raters, the structure of data distribution between two groups is clearly different. Note that the assigned scales and percentages are only the suggestion to compensate for the subjective evaluation. Nonnative teachers could modify our suggestion to the real practice depending on the situation, level of test-takers, or gender.

4.3. Pause Duration

Results of the RM ANOVA showed a significant effect of group on the pause duration between the two iPs, $F(1, 125) = 67.545$, $p < .001$. Tukey's tests ($p < .05$) revealed that the pause duration was longer for the Korean raters, and shorter for the native raters (see Figure 5). Out of the various parameters for prosody, duration of pause was usually associated with fluent English pronunciation because its degree affected listener's determination on foreign accent (Alberchtsten, Henriksen, & Faerch, 1980; Trofimovich & Baker, 2007). Pause could be interpreted as universal difficulty for second language learners, not L1 interference, because it reflected processing or memory constraints unique to L2 speech (Schachter, Christenfel, Ravina, & Bilous, 1991) as well as developmental process of L2 (Kang & Ahn, 2003). It means that pause could be the major factor to decide the L2 development universally. Following figures are two group's mean values with standard errors for three acoustic parameters:

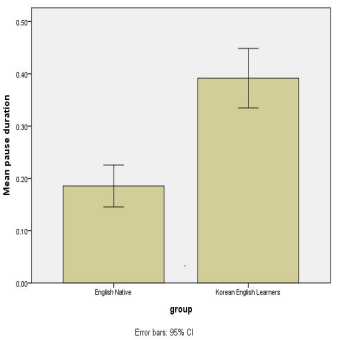
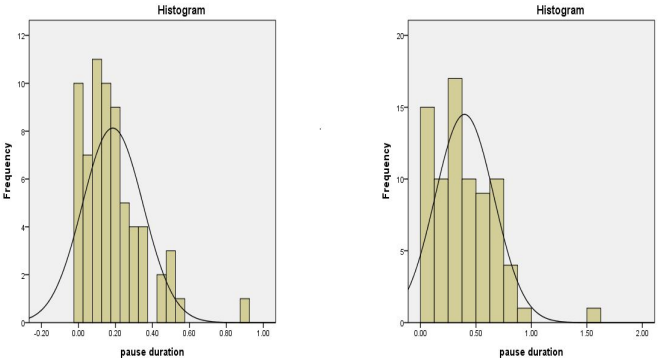


Figure 5. Pause duration of good pronunciation evaluated by natives (left) and nonnatives (right)

It is argued that pause duration is important factor to decide the fluent pronunciation. For more depth analysis, the relationship between pause duration and good speech evaluation is investigated as follows:



(a) native English raters (b) Korean L2 raters

Figure 6. Two histograms of pause duration evaluated by both groups.

Good speech of 30% evaluated by native English speakers is distributed within 0.10 seconds and 60% within 0.19 seconds. On the contrary, good speech of 30% evaluated by Korean English learners exists within 0.21 seconds and 60% within 0.45 seconds. Even though the pause duration is a good indicator of fluent L2 speaking universally (Trofimovich & Baker, 2007), the raters' L1 background influences on goodness judgment. It is clear that the frequent distribution of data between two groups is different. Nonnative teachers should be aware of L1 effect in that pause duration could be the solid indicator to the objective assessment.

5. Implication and Conclusion

Good teaching for English pronunciation by Korean teachers should go well with fair evaluation. L2 Teachers, regardless of L1 nativeness, have to assess learners' linguistic ability, progress, and achievement with fairness and reliability. Considering that English pronunciation evaluation is based on

learner's performance ability, an objective method is needed to be included in the fluency decision for the learner's speaking ability.

The results eventually confirm that L2 fluent evaluation could be affected by L1 background language. Thus, native English raters hold following speech features: a wider F0 range (Backman 1979, Jenner 1976, Willems, 1982), a faster speech rate (Guion, Flege, & Loftin, 2000; Guion, Flege, Liu, Yeni-Komshian,& Grace, 2000), and shorter pause duration (Munro & Derwing, 2006; Trofimovich & Baker, 2007). On the other hand, Korean English raters follows the patterns of a smaller F0 range and a longer duration through the whole sentence, and a longer duration of the pause.

As a kind of objective method for the prosody evaluation, a objective grading system should be proposed based on the three fluency parameters such as F0 range, speech rate, and pause duration that prove to be salient parameters in L2 acquisition. This analysis plays a very critical role considering the fluency evaluation of L2 learner's pronunciation proficiency. It doesn't necessarily mean, however, that this method is of no help since we definitely need a way to assess L2 learner's oral proficiency.

Now it is a well-known fact that the assessment of English speaking classes which have been implemented in a few years includes a pronunciation test. Since the speaking test is easier to get blame for its subjectivity in evaluation, this approach to fluency evaluation suggests that nonnative English raters should get trained for their reliable evaluation.

References

Adams, C., & Munro, R. (1978). In search of the acoustic correlates of stress: Fundamental frequency, amplitude, and duration in the connected utterances of some native and nonnative speakers of English. *Phonetica, 35,* 125-156.

Alberchtsen, D., Henriksen, B., & Faerch, C. (1980). Native speaker reactions to learners' spoken interlanguage. *Language Learning, 30,* 365-396.

Aoyama, K., Guion, S. A., Flege J. E., Tsuneo Y., & Akahane-Yamada R. (2008). The first years in an L2-speaking environment: a comparison of Japanese

children and adults learning American English. *IRAL, 46,* 61-90.

Backman, N. E. (1979). Intonation errors in second language pronunciation of eight Spanish speaking adults learning English. *Interlanguage Studies Bulletin, 4*(2), 239-266.

Bachman, L. F. (2004). *Statistical analyses for language assessment.* Cambridge, UK: Cambridge University Press.

Choi, S-K, (2011). The reliability study on the writing assessment used by Rasch models. *Reading Studies, 25,* 415-445.

Cruttenden, A. (1997). *Intonation*. (2nd edition). Cambridge: Cambridge University Press.

Derwing, T., & Munro, M. (1997). Accent, intelligibility, and comprehensibility. *Studies in Second Language Acquisition, 19,* 1-15.

Derwing, T., Thompson, R., & Munro, M. (2006). English communication and fluency development in Mandarin and Slavic speakers. *System, 34,* 183-193.

Flege, J. E. (2006). Phonetics approximation in second language acquisition. *Language Learning, 30,* 117-134.

Flege, J. E., & Fletcher, K. L. (1992). Talker and listener effects on degree of perceived foreign accent. *Journal of Acoustical Society of America, 91,* 370-389.

Guion, S. G., Flege, J. E., & Loftin, J. D. (2000). The effect of L1 use on pronunciation in Quichua-Spanish bilinguals. *Journal of Phonetics, 28,* 27-42.

Guion, S. G., Flege, J. E., Liu S. H., Yeni-Komshian, & Grace H. (2000). Age of learning effects on the duration of sentences produced in a second language. *Applied Psycholinguistics, 21,* 205-228.

Johns-Lewis, C. (1986). Prosodic differentiation of discourse modes. In C. Johns-Lewis (Ed.), *Intonation in discourse* (pp. 199-219). London: Croom Helm.

Jun, S. (2005). *Prosodic typology: The phonology of intonation and phrasing.* London: Oxford University Press.

Kang, S-H. (2013). The study on Korean raters' characteristics for Korean English oral performance. *Studies in Linguistics, 26,* 1-21.

Kang, S-H., & Ahn, H-K. (2010). The automatic measurement of prosody in English pronunciation test. *Journal of Applied Linguistics, 26*(4), 121-150.

Kim, J-K. (2006). A study on the interviewer discourse in an oral proficiency interview test. *Studies in English Language and Literature, 49*(2), 61-81.

Kim, Y-H. (2009). An investigation into native and non-native teachers' judgements of oral English performance: A mixed methods approach. *Language teaching, 26*(2), 187-217.

Lee, Y-G. (2003). *A study on the form and the function of Korean intonation.* Unpublished Master Thesis. Seoul National University.

Lennon, P. (2002). Investigating fluency in EFL: A quantitative approach. *Language Learning, 40*, 387–417.

Lincare, J. M. (2008). *A user's guide to facets: Rasch-model computer programs.* www. winsteps. com

Luoma, S. (2004). *Assessing speaking.* Cambridge: Cambridge University Press.

Mennen, I. (2006). Phonetic and phonological influences in non-native intonation: An overview for language teachers. *Queen Margaret University College speech science research center working paper. 9*, 1-18.

Munro, M., & Derwing, T. (1996). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 45*(1), 73-97.

Munro, M., & Derwing, T. (2008). Segmental acquisition in adult ESL learners: A longitudinal study of vowel production. *Language Learning, 58,* 479-502.

Myford, C., & Wolfe, W. (2004). Detecting and measuring rater effects using many-facet Rash measurement: Part II. *Journal of Applied Measurement, 5*(2), 189-227.

Schachter, S., Christenfeld, N., Ravina, B., & Bilous, F. (1991). Speech disfluency and the structure of knowledge. *Journal of Personality and Social Psychology, 60,* 362-367.

Scherer, K. R. (2000). A cross-cultural investigation of emotion inferences from voice and speech: Implications for speech technology. In B. Yuan, T. Huang, & X. Tang (Eds.), *Proceedings of the sixth international conference on spoken language processing* 2 (pp.379-382). Beijing: China Military Friendship Publish.

Segalowitz, N. (2010). *Cognitive bases of second language fluency.* New York: Routledge.

Shin, D. I. (2002). The development of rating system for High school English

writing and speaking using Rasch model. *English Education, 57,* 469-499.

Sluijter, A. C., & Heuven, V. V. (1996). Acoustic correlates of linguistic stress and accent in Dutch and American English. *International Conf. on Spoken Language Process, 2,* 630-633.

Trofimovich, P., & Baker, W. (2006). Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition, 28*(1), 1-30.

Willems, N. J. (1982). *English intonation from a Dutch point of view*. Dordrecht: Foris Publication.

Winke, P., Gass, S., & Myford, C. (2012). Raters' L2 background a potential source of bias in rating oral performance. *Language Testing, 30(*2), 231-252.

Wright, D., & Linacre, M. (1994). Reasonable mean-square fit values. *Rasch measurement transactions SIG*, *8,* 370.

Zhang, Y., & Elder, C. (2011). Judgements of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing, 28*(1), 31-50.

Seokhan Kang
#405, 10-1 Dong
Institute of Foreign Language Education
Seoul National University
599, Gwanak-ro, Gwankak-gu
Seoul, 151-748, Korea
Phone: 02-880-7616, 010-9120-2433
Email: kangs45@snu.ac.kr
Website: seokhan.wordpress.com

Received on September 25, 2013
Revised version received on November 30, 2013
Accepted on December 10, 2013