Native and non-native speakers' perceptions of fluency in L2 speech

Jinyoung Jo
(Seoul National University)

Jo, Jinyoung. 2015. Native and non-native speakers' perceptions of fluency in L2 speech. *SNU Working Papers in English Linguistics and Language 13, 40-62.* The purpose of this study is to explore the relationship between objectively measured acoustic properties of speech on the one hand and fluency as rated by native and non-native raters on the other hand. Acoustic measures of fluency were calculated for each speech sample recorded by eleven Korean EFL learners, and two native English speakers and two native Korean speakers rated these speech fragments on fluency. The regression analysis demonstrated that while native and non-native speakers agreed on the acoustic properties they considered important in assessing fluency, i.e., mean length of runs and average pause time, the latter group were found to rely more on those objectively measured acoustic features. An examination of the comments made by both groups of raters further revealed that only native speakers regarded intelligibility as an important factor in judging fluency. The findings of this study have critical implications in the field of speaking tests; they suggest not only that it is important to consider what characterizes a fluent speech but also that raters' L1 background must be taken into consideration in assessing fluency of an L2 learner's speech. **(Seoul National University)**

Keywords: Speaking tests, acoustic measures of fluency, perceived fluency, L2 speech, rater L1 background

1. Introduction

The rating of language proficiency is often performed by human raters. Rater variables have been major concerns in language testing for ensuring reliability and validity of a test. Raters' language background is one important factor that can influence ratings in second language speaking tests. The purpose of this study is to explore the influence of raters' L1 background on ratings of perceived fluency. More specifically, this study investigates the relationship between objectively measured acoustic properties of speech on the one hand, and fluency as rated by native and non-native raters on the other hand.

A great deal of research has investigated fluency in L2 speech in terms of native speakers' perceptions. Most of them explored which acoustic measures of fluency can predict perceived fluency in L2 speech. However, little has been known about how raters' L1 background can play a role in this aspect. In this paper, regression analysis will reveal the difference in native (i.e., L1 English) and non-native (i.e., L1 Korean) speakers' fluency ratings, with respect to which specific acoustic properties are more strongly related to perceived fluency.

2. Previous studies

In this section, I will review findings from previous research in three aspects: (1) Influence of raters' L1 background on ratings in speaking tests, (2) relationship between acoustic measures and perceived fluency and (3) relationship between fluency and intelligibility.

2.1 Influence of raters' L1 background on ratings in speaking tests

Numerous studies have investigated influence of raters' language background on ratings in speaking tests. Findings from these studies do not point to a single conclusion, as they were conducted in different contexts as regards to, for example, proficiency level of test takers and particular features of tests. The most widely examined areas are influence of raters' L1 background on severity and consistency of ratings. With regard to severity of ratings, some studies have found that non-native speakers were harsher than their native speaker counterparts (e.g., Fayer and Krasinski, 1987; Yu, 2010), while others demonstrated the opposite tendency (e.g., Barnwell, 1989; Hill, 1996). A number of studies have shown that non-native raters are more consistent than native raters (e.g., Brown, 1995; Hill, 1996), while others demonstrate the opposite outcomes (e.g., Shi, 2001). Yet another body of research found no significant influence of raters' L1 background on ratings in speaking tests (e.g., Kim, 2009; Wei & Llosa, 2015; Zhang & Elder, 2011). As various dimensions of proficiency were assessed and different target language was investigated in different studies, no consensus has been reached regarding how raters' L1 background may influence rating behaviors.

In addition to severity and consistency in ratings, many studies have also examined how native and non-native speakers differ from each other in assessing various constructs of oral proficiency. The findings revealed that they evaluated the test takers' performances similarly in some respects and differently in others. Fayer and Krasinski (1987) examined how English native speakers and Spanish native speakers differed in evaluating English oral proficiency of Puerto Rican test takers. The results demonstrated that while both groups gave similar scores on intelligibility, Spanish raters were more severe in assessing linguistic forms of the test takers' speech. Zhang and Elder (2011) investigated how Chinese native speakers' ratings differed from those of English native speakers' in rating Chinese test takers' oral proficiency. Results showed that while there was no significant difference in raters' holistic judgments of the speech samples, their comments revealed that they differed with respect to which constructs of oral proficiency they considered relatively more important. More specifically, non-native raters were found to weigh vocabulary and general linguistic resources more than other criteria, while native raters mentioned that they considered interaction and compensation strategies as more important factors of proficiency. Also, native speakers tended to make more comments on fluency of the test takers' speech than non-native speakers, but this difference was not statistically significant. A more clear result was given in Gui (2012). It was reported in this study that native English and Chinese speakers had different opinions about fluency of Chinese students' speech in English; American raters were found to provide harsher comments on the delivery of the students' speech.

A great deal of research has investigated different rating patterns shown by English native speakers and Korean native speakers when they assess Korean students' L2 speech in English. For example, in Kim's study (2006), it was found that native English speakers differed from native Korean speakers in several analytic ratings including rate of speech, which is a measure of fluency, as well as organization and task fulfillment. According to Kim (2009), native and non-native raters exhibited no difference in severity or consistency of their ratings; however, they did diverge on which specific criteria they referred to when making judgments on students' performance. Native raters' comments were found to be more detailed than those of the non-native raters in several areas including pronunciation. Finally, Yu (2010) found that nonnative raters awarded lower scores for fluency than the native raters did across all levels. However, this difference was only significant for test takers with a lower proficiency level; for more proficient test takers, the two rater groups did not show a statistically significant difference in scoring.

In sum, various studies suggested different, sometimes opposite, findings on native and non-native speakers' rating behaviors. An extensive discussion in the literature involves different rating behaviors between L1 English raters and L1 Korean raters in assessing fluency of speech produced by Korean EFL learners. This study further investigates this issue by incorporating objectively measured acoustic properties of speech on the one hand, and fluency as rated by native and non-native raters on the other hand.

2.2 Relationship between acoustic measures and perceived fluency

In this subsection, I will review the concept of fluency as recognized in previous studies and present the results of experiments that investigated the relationship between acoustic measures of fluency and perceived fluency. Fluency may be defined as an “automatic procedural skill” (Schmidt, 1992) that includes “rapid, smooth, accurate, lucid, and efficient translation of thought or communicative intention into language” (Lennon, 2000, p. 26). According to Segalowitz (2010), fluency is assumed to have three facets: cognitive fluency, utterance fluency and perceived fluency. First, cognitive fluency refers to the speaker’s ability to plan and implement speech by integrating the cognitive mechanisms underlying speech execution. Second, utterance fluency refers to the fluency that can be measured by objective acoustic properties of an utterance. Utterance fluency consists of three components, i.e., speed fluency, breakdown fluency and repair fluency. Speed fluency is concerned with the speed of speech, breakdown fluency with the number and length of pauses and repair fluency is related to the number of false starts, corrections and repetitions. Lastly, perceived fluency is the fluency judged subjectively by raters. Relevant to the present study are utterance fluency and perceived fluency; in particular, this study investigates the relationship between the two facets of fluency by exploring which specific measures of utterance fluency is more strongly related to perceived fluency judged by two groups of raters with different L1 backgrounds.

Numerous studies have examined which acoustic properties are more strongly related to perceived fluency. It should be noted that as different studies adopted different acoustic measures of fluency with varying target languages, their findings are somewhat inconclusive with regard to which features of utterance fluency can best predict perceived fluency. In the rest of this subsection, major findings from two recent experiments are reported. First, Pinget, Bosker, Quené and de Jong (2014) investigated the relationship between objectively measured temporal properties of speech and fluency rated by native speakers. The target language was Dutch and test takers' mother tongue was either Turkish or English. The results revealed that six acoustic measures adopted were able to explain a large part of the variance in fluency ratings. Especially, a measure of speed fluency was found to be the best predictor of perceived fluency, followed by length of pauses. Second, Préfontaine, Kormos and Johnson (2015) posited four acoustic measures which were expected to predict fluency ratings judged by French native speakers. In the study, mean length of runs, a measure of both speed and breakdown fluency, was found to be the best predictor of perceived fluency. While it is hard to make a generalization from only these two studies with respect to which measures of utterance fluency is more strongly related to perceived fluency, it seems that both speed and breakdown fluency is considered important by native speakers in judging fluency of L2 speech.

In summary, native speaker raters were found to regard speed and breakdown fluency as important criteria in assessing fluency of test takers, though results may vary depending on the target language and which specific acoustic measures were included as potential predictors. It should be noted that in this line of research in which the relationship between utterance fluency and perceived fluency is investigated, perceived fluency is rated by only native speakers of the target language in most cases. Considering that raters' L1 background may influence ratings in speaking tests as suggested in the previous section, native and non-native raters are likely to differ in terms of how they rate fluency of L2 speech and especially with regard to which acoustic measures their judgments are based on. This study further addresses this issue, by comparing the two groups of raters' ratings in relation to objectively measured acoustic properties of speech samples.

2.3 Relationship between fluency and intelligibility

Although not the main interest of this study, it is necessary to take a look at the findings from previous research on relationship between fluency and intelligibility, since an examination of the comments made by the two groups of raters in the present experiment has an implication on this issue. As will be shown in Section 4, raters' L1 may also play a role in dynamics of the relationship between fluency and intelligibility.

Findings from previous studies indicated that fluency and intelligibility may be weakly related, where intelligibility is operationalized as "how accurately listeners are able to identify spoken language relative to an L2 speaker's intended utterance" (Thompson, 2015, preprint version). For example, Derwing and Munro (1997) found that speech rate as well as prosody was significantly correlated with the intelligibility scores of only a small subset (8%) of listeners in their study. In the same study, 15% of the raters indicated that they were consciously aware that several features associated with fluency had affected their comprehensibility judgments.

In this paper, qualitative comments provided by two groups of raters will reveal that a rater's L1 background may have influenced whether he or she perceived fluency and intelligibility as two related constructs of oral proficiency.

3. Methods

3.1 Acoustic measures of fluency

Different studies have adopted a wide variety of acoustic measures of fluency. In the present study, four acoustic measures were selected to be examined, adopted from Préfontaine et al. (2015). Articulation rate was chosen as a measure of speed fluency, and pause frequency and average pause time as measures of breakdown fluency. In addition, mean length of run represents both speed and breakdown fluency.[1] These variables

[1] I did not include the number of filled pauses as a measure of fluency. It is argued that

were operationalized as follows:

1. Articulation rate (AR): The total number of syllables divided by the total phonation time in seconds. (Phonation time refers to the duration of an entire speech excluding pauses of 0.25 seconds or longer.[2])
2. Mean length of runs (MLR): The total number of syllables divided by the number of utterances between pauses.
3. Pause frequency (PF): The total number of pauses divided by the total duration in seconds.
4. Average pause time (APT): The total duration of all pauses divided by the number of pauses.

Acoustic properties of speech samples were measured by using Praat (Boersma & Weenick, 2015) and a Praat script (De Jong & Wempe, 2009).

3.2 Testing Procedures

3.2.1 Participants

Eleven Korean undergraduate students studying at Seoul National University (SNU) were recruited. Each test taker's level of English proficiency varied from intermediate-low to high. A questionnaire distributed after the experiment revealed that most of them have learned English for more than ten years. Five students reported they have experiences of studying in countries where English is the main spoken language. Their learning experiences did not exceed two years. All

how many filled pauses a speaker makes is not an accurate parameter of judging fluency in L2 speech, as native speakers also produce a lot of filled pauses.

[2] Only pauses of 0.25 seconds or longer were intended to be detected by Praat. The default setting of Praat scripts was arranged as such; also, this cut-off point of pause length is the most widely adopted criterion for defining pause in the literature (De Jong & Bosker, 2013; Pinget et al., 2014; Préfontaine, 2015).

participants took part in the experiment on a voluntary basis.

3.2.2 Speaking tasks

Participants were asked to provide answers in English to three questions. The topics were intended to be easy and familiar to students, with the content assumed to occur frequently in daily conversations. It was based on the idea that topics that are too difficult to handle may hinder students' demonstration of fluency in L2 speech. Also, the questions were intended to have an equal degree of familiarity and difficulty. The following are three questions given to the students:

1. Do you prefer to travel by yourself or with others such as family and friends? Give specific reasons and details to support your answer.
2. Do you perform better when you are competing or when you are collaborating with others? Give specific reasons and details to support your answer.
3. Describe a memorable movie that you watched lately. Explain why you think this movie is impressive or interesting.

3.2.3 Recording procedures

Recording was done in a recording lab in SNU using Praat at a sampling rate of 44,100 Hz. The task takers were given 30 seconds to prepare their answers and 60 seconds to speak for each question. For each speech sample of approximately one minute, 20-seconds recordings were extracted from the middle of the original one. In this way, a total of 33 speech fragments (11 speakers × 3 tasks) were obtained.

3.3 Rating procedures

3.3.1 Raters

Two native raters (i.e., L1 English) and two non-native raters (i.e., L1 Korean) were employed. The two American raters had been living in Korea for less than three months and spoke minimal Korean. The two non-native raters had a high proficiency level of English. One was an undergraduate student at SNU, the other a graduate student.

3.3.2 Procedures

Raters were given detailed instructions and scoring guidelines before they began rating. The instructions included the goal of the study and definition of fluency as noted in Section 2.2 (Lennon, 2000). Adapted from Préfontaine et al. (2015), raters were informed that a fluent speaker can express himself or herself spontaneously at length with a natural colloquial flow at reasonable speed and with few pauses. They were also told that a less fluent speaker can manage very short isolated utterances, with much pausing to search for expressions. Importantly, they were instructed that fluency is not the same construct as accentedness, as accent is mostly concerned with pronunciation and intonation.

Raters were asked to rate each speech sample's fluency based on a scale from 1 to 9 (1: not fluent at all, 9: very fluent). After they finished ratings, they were also asked to freely express what factors have affected their judgments. As was the case in previous studies (e.g., Gui, 2012; Kim, 2009; Zhang & Elder, 2011), raters' qualitative comments were expected to provide valuable information as regards to which acoustic properties of speech were considered crucial in assessing fluency.

4. Results

In this section, results of the experiment are presented in two steps. First, a descriptive analysis on the objectively measured acoustic properties of the speech samples and the scores given by the raters will be presented. Second, I will demonstrate the results of a regression analysis which indicated that native and non-native speakers' rating behaviors differed in terms of the relationship between acoustic measures and perceived fluency ratings.

First, I computed the descriptive statistics for the four acoustic properties of 33 speech samples. The results are presented in Table 1.

Table 1. *Acoustic properties of the test takers' speech (AR: articulation rate, MLR: mean length of run, PF: pause frequency, APT: average pause time)*

	Task1 Mean (SD)	*Task 2* Mean (SD)	*Task 3* Mean (SD)	*Task 1, 2, 3* Mean (SD)
AR	3.38 (0.24)	3.46 (0.50)	3.43 (0.41)	3.43 (0.39)
MLR	6.38 (3.24)	5.63 (2.03)	5.45 (1.49)	5.82 (2.33)
PF	0.40 (0.11)	0.43 (0.09)	0.41 (0.06)	0.41 (0.08)
APT	0.92 (0.36)	0.80 (0.26)	0.85 (0.32)	0.86 (0.31)

Second, I calculated descriptive statistics for the native and non-native speakers' ratings of perceived fluency. The statistics are provided in Table 2.

It was found that non-native speakers awarded significantly lower scores than their native speaker counterparts, indicated by the two groups' mean scores of 3.80 and 4.55, respectively (p < 0.05). That is, non-native speakers were harsher in their ratings than native speakers in general.

Table 2 *Ratings of perceived fluency by native and non-native speakers*

	Native rater Mean (SD)	*Non-native rater* Mean (SD)
Task 1	4.82 (2.44)	3.59 (2.09)
Task 2	4.59 (2.63)	4.00 (2.45)
Task 3	4.23 (2.09)	3.82 (2.30)
Task 1, 2, 3	4.55 (2.48)	3.80 (2.25)

The next step was to perform a linear regression analysis, with perceived fluency ratings as the dependent variable and utterance fluency measures as independent variables. All regression analyses in the present study were performed in R (R Development Core Team, 2015). The coefficients are presented in Table 3 below, for native and non-native speaker raters, respectively.

Table 3. *Correlations of acoustic measures with perceived fluency by raters' L1 background. (* p < .05, ** p < .01, *** p < .001)*

	Fluency ratings (native)	*Fluency ratings (non-native)*
AR	0.162	0.373 **
MLR	0.357 **	0.450 ***
PF	- 0.208	- 0.223
APT	- 0.294 *	- 0.565 ***

It was discovered that only a subset of acoustic measures was significantly correlated with fluency ratings. More importantly, native and non-native raters' ratings differed from each other regarding which measures of utterance fluency they were correlated with. For native speakers’ ratings, only MLR and APT were significantly correlated with perceived fluency. The coefficient for MLR was 0.357, indicating that the greater the number of syllables contained in an utterance, the higher the fluency score. The coefficient for APT was - 0.294, meaning that speakers with longer average pause times were judged to be less fluent. For non-native speakers’ ratings, three acoustic measures, i.e., AR, MLR and APT, were significantly correlated with perceived fluency. It should be noted that AR was significantly correlated only with non-native speakers’ ratings. Moreover, the coefficients of MLR and APT, which were significantly correlated with ratings by both groups, were higher for non-native speakers.

For both groups, stepwise regressions were implemented to find out which acoustic measures can best predict variance of fluency ratings. For native speakers’ ratings, only APT and MLR were found to explain the variance of the scores. A linear regression was conducted with perceived fluency ratings as the dependent variable and APT and MLR as independent variables, in order to find out how much of the variance can be explained with these two factors. The results are shown in Table 4.

Table 4. *Effects of utterance fluency measures on perceived fluency (native speakers)*

	Estimate	*Standard Error*	*Pr(>|z|)*
(Intercept)	4.1956	1.1955	0.000835 ***
APT	-1.7205	0.9087	0.062910
MLR	0.3132	0.1210	0.011946 *

Adjusted $R^2 = 0.148$

The adjusted R^2 of the model with the two measures as predictors indicated that only 14.8% of the variance in fluency ratings may be accounted for by these two factors. Note that MLR had a significant effect on fluency ratings and APT had a marginally significant effect.

The same method was used for non-native speakers' ratings. A regression analysis demonstrated that as was the case in native speakers' ratings, the final model included only APT and MLR as accounting for the variance in non-native speakers' ratings. A linear regression was conducted with fluency ratings as the dependent variable and APT and MLR as independent variables. Table 5 below shows the results.

Table 5. *Effects of utterance fluency measures on perceived fluency (non-native speakers)*

	Estimate	*Standard Error*	*Pr(>|z|)*
(Intercept)	4.9435	0.9462	2.10e-06 ***
APT	-3.5653	0.7192	5.69e-06 ***
MLR	0.3284	0.0957	0.00107 **

Adjusted $R^2 = 0.408$

The adjusted R^2 of the model with APT and MLR as predictors indicated that approximately 40.8% of the variance in fluency ratings can be explained by these two factors. The adjusted R^2 of this model was much higher than that of native speakers' ratings. Both APT and MLR were found to account for variance of the perceived fluency at a significant level.

5. Discussion

In this section, the results of the experiment will be reviewed, focusing on the difference between L1 English raters and L1 Korean raters with respect to the relationship between their fluency ratings and objectively measured acoustic properties of speech samples. In addition, qualitative comments made by the raters will be reported, which may partly explain the lower explanatory power of the acoustic measures for the native speakers' ratings than for non-native speakers'.

The results of the experiment demonstrated several important findings. First, recall that scores given by native speakers were in general higher than those of non-native speakers. It shows that non-native speakers were harsher in assessing fluency of L2 learners' speech. This result is in line with Yu (2010) but conflicting with Gui (2012) (see Section 2.1.). However, in order to obtain a more reliable result, it is necessary to recruit a greater number of test takers and raters.

Crucially, a regression analysis revealed that for both native and non-native raters' perceived fluency scores, only APT and MLR were found to be good at explaining the variance of the scores; AR and PF had little explanatory power. The role of APT and MLR in predicting perceived fluency is well recognized in the literature. Previous research suggested that mean length of pauses was negatively correlated with fluency ratings (Bosker et al., 2013; Pinget et al., 2014), i.e., the longer the pauses, the less fluent a speech is perceived, though a few studies argued that they were not related (Kormos & Denes, 2004). The present study confirmed that the raters, both native and non-native speakers, considered APT an important factor in judging fluency. In addition, the results revealed that MLR can also be a good predictor in accounting for the variance in perceived fluency ratings. Previous studies have also shown that MLR contributes to explaining the variance in scores of perceived fluency (Préfontaine et al., 2015; Towell, 2002). The present study confirmed that MLR is regarded as a critical factor in determining fluency of L2 speech.

In contrast, it was discovered that AR and PF did not contribute much to explaining the variance of scores in fluency ratings by both rater groups. While AR was positively correlated with fluency ratings of non-native raters, it was not added as a predictor of fluency ratings.

This result is somewhat unexpected, since a number of previous studies have suggested that AR is an important variable in predicting perceptions of fluency (Préfontaine et al., 2015; Towell et al., 1996). The same pattern is exhibited for PF. While PF in this study was not even significantly correlated with fluency ratings by the two groups, previous research has argued for its explanatory power in predicting perceptions of fluency. For example, Derwing et al. (2004) found a significant correlation between PF and perceived fluency. Similarly, Pinget et al. (2014) discovered that the number of silent pauses per second spoken time, a comparable construct to PF in the present study, was able to explain some proportion of variance of fluency ratings.

With regard to the reason why AR and PF were not good predictors of perceived fluency, two possible explanations can be provided. One speculation is that the number of speech fragments and raters were too small. Only 11 students participated in the experiment, yielding 33 speech samples, and they were rated by four raters, two of them being native English speakers and the other two Korean. In comparison, in other studies investigating the relationship between utterance and perceived fluency, a lot more participants were recruited. For example, in Pinget et al. (2014), 40 learners of French took part in the experiment and their speech fluency was evaluated by 11 native speakers. Similarly, in Préfontaine et al. (2015), a total of 90 speech fragments were assessed by 20 native speakers. Another possible explanation is that what is more important than PF is the location of pauses and their distribution, as suggested in Ejzenberb (2000) and Riggenbach (1991). This idea came from Préfontaine et al. (2015), in which a similar outcome to the present study was obtained; that is, PF accounted for only a small proportion of the variance in scores of perceived fluency. Crucially, support for this hypothesis also comes from qualitative comments made by one of the raters in the present experiment. Our first non-native speaker mentioned after rating procedures that location of pauses, along with their length, affected his ratings. This suggests a possibility that raters may have put more emphasis on location of pauses than their frequencies. More research on this issue is necessary.

As discussed above, as native and non-native raters were shown to consider the same acoustic measures as important factors, i.e., APT and MLR, one might argue that they share the same rating behaviors. However, a closer look at the results of the regression analysis sheds light on the differences between the two groups of raters. More specifically, it can be argued from the findings that non-native speakers relied more heavily on acoustic properties of speech than the native speakers did. For native speakers' ratings, the adjusted R^2 of the model with APT and MLR added as predictors of fluency showed that only 14.8% of the variance in native raters' fluency ratings may be explained based on the acoustic measures, while the adjusted R^2 for the model with the same predictors was much higher for non-native speaker group, indicating that about 40.8% of the variance of their ratings can be explained. Therefore, we can draw a conclusion that even though APT and MLR were good predictors of perceived fluency in both groups, their explanatory power was greater in the non-native raters' fluency ratings. It means that non-native speakers were more likely to rely on utterance fluency when subjectively judging fluency of a speech sample. This argument is largely supported by qualitative comments provided by each rater after they finished rating procedures. They were asked to freely express their ideas on what factors have affected their judgment. Their comments revealed that while native speakers put more emphasis on the intelligibility of the speech as one important factor in determining a fluent speech, their non-native counterparts made no reference to such criterion. Table 6 below summarizes the raters' comments.

Table 6. *Qualitative comments on what factors the raters considered important. (O = was commented to be important, X = was commented to be unimportant, blank = was not mentioned; additional information*

about the raters' comments was given below the notations O and X.)

	Native 1	Native 2	Non-native 1	Non-native 2
Speed	O Speed mattered, but doesn't have to be fast as long as the speaker sounded comfortable.	O Even if the speech was understandable, speed also mattered (i.e., more fluent when spoken quickly and eloquently)		O
Length of pauses		O	O	O
Filled pauses	X	O		O
Repetition			O	O
Intelligibility	O "As long as I got the meaning…"	O "If I can understand what they are trying to say…"		
Vocabulary	O Ability to use chunks or idioms	O	O	
Others			Location of pauses, grammar, corrections	Number of pauses, number of sentences, completeness of sentences

In Table 5, “O” indicates the features of proficiency that each rater claimed to have affected their judgment, while “X” denotes those that were claimed not to matter much in rating fluency. Note that while non-native raters did not make any comments about how they thought about the intelligibility, both of the two native raters mentioned that they tended to assign higher scores for a speaker as long as they could understand the meaning conveyed by the speech. Such comments are in line with the quantitative results of this study in that native raters were less likely base their judgments on acoustic measures of fluency. Recall that a weak relationship was found between fluency and intelligibility in previous research (see Section 2.3.). It can be argued from the findings of the present study that a rater’s L1 background may influence how intelligibility of a speech might affect fluency perceived by the rater.

Several other remarks made by the raters can explain the results of the experiment. The raters’ comments on speed of speech seem to give an explanation for why AR was not a good predictor of perceived fluency. It can be said that only two of the raters, i.e., the second native rater and the second non-native rater, found this factor significantly important in assessing fluency. The first native rater did take speed into consideration, but she did not think that faster speech meant more fluent speech. Instead, she mentioned that as long as the test taker spoke with a reasonable speed and sounded comfortable, she regarded the speaker as fluent. This may partly explain why AR was not a strong predictor of fluency ratings in this study, showing a conflicting result with previous studies. Regarding length of pauses, recall that APT contributed to explaining the variance of fluency in both groups’ ratings. This is also reflected in the raters’ comments, as three out of the four raters stated that length of pauses had affected their ratings. In terms of PF, only one rater, i.e., the second non-native speaker, made a comment on the number of pauses. This explains why PF was not a good predictor that accounts for variance in fluency ratings. Interestingly, none of the raters made comments concerned with MLR, which was actually proven to be a significant factor in quantitative analysis. My speculation is that while raters were unknowingly affected by MLR of a speech, they are not aware of the fact that MLR could influence their judgment. In summary, raters' qualitative remarks revealed that the features of speech that were found to be crucial in predicting perceived fluency in quantitative analysis corresponded to those in qualitative analysis to a great degree.

6. Conclusion

This study explored the relationship between objectively measured acoustic properties of speech on the one hand and fluency as rated by native and non-native raters on the other hand. Acoustic measures of fluency were calculated for each speech sample recorded by Korean EFL learners, and native English speakers and native Korean speakers were asked to rate these speech fragments on fluency. The regression analysis demonstrated that while native and non-native speakers agreed on the acoustic properties they considered important in assessing fluency, the latter group were found to rely more on those objectively measured acoustic features. Examination of the statements made by both groups of raters further revealed that only native speakers regarded intelligibility as an important factor in judging fluency.

The findings of this study bring about important implications to the field of speaking tests. As automated measures are to be used in speaking tests, the present study suggests not only that it is important to consider what characterizes a fluent speech but also that raters' L1 background must be taken into consideration in this respect. However, several issues remain unresolved. First, the number of test takers and raters should be augmented in future studies. In this way, we can further investigate the reason why AR and PF were found to be no good predictors of perceived fluency; that is, we can determine if such result was due to a limited number of speech samples and raters or there were some other causes underlying the two measures' weak explanatory power. Second, as was revealed by the raters' comments, it would be of great interest to examine the effects of other factors of utterance fluency on perceived fluency. For example, the comments indicate that the raters had diverging opinions about the effects of filled pauses, repetitions and vocabularies on fluency ratings. In future studies, these factors can be included as potential predictors in explaining the variance of perceived fluency.

References

Barnwell, D. (1989). 'Naive' native speakers and judgements of oral proficiency in Spanish. *Language Testing, 6*(2), 152-163.

Boersma, P., & Weenink, D. (2015). Praat: doing phonetics by computer (Version 5.4.08) [Computer software]. Retrieved www.praat.org

Bosker, H., Pinget, A., Quené, H., Sanders, T., & De Jong, N. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing, 30*, 159–175.

Brown, A. (1995). The effect of rater variables in the development of an occupation specific language performance test. *Language Testing, 12*(1), 1-15.

De Jong, N.H., & Bosker, H.R. (2013). Choosing a threshold for silent pauses to measure second language fluency. In R. Eklund (Ed.), *Proceedings of Disfluency in Spontaneous Speech.* (pp. 17-20). Stockholm, DiSS 2013.

De Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods, 41*, 385–390.

Derwing, T., & Munro, M. (1997). Accent, intelligibility, and comprehensibility. *Studies in Second Language Acquisition, 1*, 1-16.

Derwing, T., Rossiter, M., Munro, M., & Thomson, R. (2004). Second language fluency: Judgments on different tasks. Language Learning, 54, 655–679.

Ejzenberg, R. (2000). The juggling act of oral fluency: A psycho-sociolinguistic metaphor. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 287–313). Ann Arbor, MI: University of Michigan Press.

Fayer, J. M., & Krasinski, E. (1987). Native and nonnative judgements of intelligibility and irritation. *Language Learning, 37*(3), 313-326.

Gui, M. (2012). Exploring Differences Between Chinese and American EFL Teachers' Evaluations of Speech Performance. *Language Assessment Quarterly, 9,* 186–203.

Hill, K. (1996). Who should be the judge? The use of nonnative speakers as raters on a test of English as an international language. *Melbourne Papers in Language Testing, 5*(2), 29-50.

Kim, H. J. (2006). Rater reliability in L2 oral proficiency tests. *English Teaching, 61*(3), 105-118.

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

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System, 32,* 145–164.

Lennon, P. (2000). The lexical element in spoken second language fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 25–42). Ann Arbor, MI: The University of Michigan Press.

Pinget, A., Bosker, H., Quené. H., & De Jong, N. H. (2014). Native speakers' perceptions of fluency and accent in L2 speech. *Language Testing, 31*(3), 349-365.

Préfontaine, Y., Kormos, J., & Johnson, D. E. (2015). How do utterance measures predict raters' perceptions of fluency in French as a second language? *Language Testing.*

R Development Core Team. (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved http://www.R-project.org.

Riggenbach, H. (1991). Toward an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse Processes, 14,* 423–441.

Schmidt, R. (1992). Psychological mechanisms underlying second language fluency. *Studies in Second Language Acquisition, 14,* 357–385.

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

Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing, 18*(3), 303-325.

Thomson, R. I. (2015). Fluency. In Reed, M. & Levis, J. *The Handbook of Pronunciation*. (pp. 209-226). Hoboken, New Jersey: Wiley.

Towell, R. (2002). Relative degrees of fluency: A comparative case study of advanced learners of French. *International Review of Applied Linguistics in Language Teaching, 40*, 117–150.

Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics, 17*, 84–119.

Wei, J., & Llosa, L. Investigating Differences Between American and Indian Raters in Assessing TOEFL iBT Speaking Tasks. *Language Assessment Quarterly, 12,* 283-304.

Yu, K. A. (2010). The Effect of Raters' Language Background on English-speaking Test Ratings across Test-takers' Oral Proficiency Levels. *Korean Journal of Applied Linguistics, 26*(4), 395-419.

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

Jinyoung Jo
jinyoungjo710@gmail.com