ISSN 2005-8063
2017. 3. 31.
Vol.9 No.1
pp. 19-26

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

http://dx.doi.org/10.13064/KSSS.2017.9.1.019

The effects of pause in English speaking evaluation*

Mi-Sun Kim · Tae-Yeoub Jang**

Abstract

The main objective of this study is to investigate the influence of utterance internal pause in English speaking evaluation. To avoid possible confusion with other errors caused by segmental and prosodic inaccuracy, stem utterances with two different length obtained from a native speaker were manipulated to make a set of stimuli tokens through insertion of pauses whose length and position vary. After a total of 90 participants classified into three proficiency groups rated the stimuli, the scored data set was statistically analyzed in terms of the mixed effects model. It was confirmed that predictors such as pause length, pause position and utterance length significantly influence raters' evaluation scores. Especially, a dominating effect was found in such a way that raters gradually deducted scores in accordance with the increase of pause duration. In another experiment, a tree-based statistical learning technique was utilized to check which of the significant predictors played a more influential role than others. The findings in this paper are expected to be practically informative for both the test takers who are preparing for an English speaking test and the raters who desire to develop more objective rubric of speaking evaluation.

Keywords: pause length, pause position, L2 pause, English speaking evaluation, predictor importance

1. Introduction

This study began with a question after identifying inconsistent rating results regarding utterance internal pauses in English speaking tests. 'What do human raters mark with higher scores: utterances with extremely long pauses or utterances with improperly positioned pauses?' There were some issues that caused discrepant scores among raters on English speaking tests. One such case was an extremely long duration of pause to which some raters assigned low scores while others did not. This phenomenon led to design this experiment so that the first objective of this test was to find out the human raters' scoring tendency according to pause length. Moreover, using pauses in proper positions is highlighted in English speaking education, as it is important in meaning construction. Thus, back to the original question, analyzing the correlation of the position and length of pause was another objective of the current experiment. Verifying human raters' priority in terms of pause-related features would be beneficial for reducing the rating gap between evaluators.

Among other prosodic features, such as rhythm, intonation and stress, pause is one of the most important features in speech naturalness (Zellner, 1994) and accentedness (Kang, 2010). Previous studies have already emphasized the importance of pause length in L2 evaluation, and there is consensus among the studies that L2 learners produce longer pauses than native L1 speakers, and lower level L2 speakers have longer pauses than their high level counterparts. This phenomenon is reported by studies of L2 learners of many different nationalities.

Anderson-Hsieh & Venkatagiri (1994) compared Chinese ESL learners' pause production at two different proficiency levels and found that low level learners produced more inappropriate and longer pauses than high level learners. Pickering (1999) also reported

* This work was supported by the 2017 Research Grant from Hankuk University of Foreign Studies
** Hankuk University of Foreign Studies, tae@hufs.ac.kr, corresponding author
Received 31 January 2017; Revised 11 March 2017; Accepted 15 March 2017

similar results on Chinese L2 learners. Riazantseva (2001) investigated the relationship between L2 proficiency levels and pausing patterns of Russian learners of English. She reported that high level learners made significantly shorter pauses than their low level counterparts. Kormos & Dénes (2004) investigated the fluency of Hungarian L2 learners at two proficiency levels and reported that advanced L2 learners produced shorter pauses than those of low level counterparts. Both high level learners in Anderson-Hsieh & Venkatagiri (1994) and Riazantseva (2001) were reported to have similar pause length as those of native speakers. However, the lower level learners produced approximately 1.3 (Riazantseva, 2001) to 1.6 times (Kormos & Dénes, 2004) longer pauses than those of higher level learners in their data of spontaneous narratives, and almost 2 times (Anderson-Hsieh & Venkatagiri, 1994) longer pauses than those of higher levels in their read speech data. As a result, Kormos & Dénes (2004) purported that the mean length of silent pauses were significantly associated with the fluency judgments. According to Trofimovich & Baker (2006), in their ratings of Korean L2 learners of English, pause duration was the most influential cue to accentedness ratings compared to other features including stress timing, peak alignment, speech rate, and pause frequency. Kang (2010) also mentioned the influence of pause duration in fluency judgment. According to her study of 11 international teaching assistants of different nationalities, abnormal boundary pauses within clauses, especially pauses over 0.8 seconds, appeared to be the strongest cue on ratings of high level L2 learners' accentedness. Munro *et al*. (2006) and Rossiter (2009) also reported similar results from the rating data of other nationalities.

There are two major restrictions in the previous research on L2 utterances with pause and their fluency ratings. First, diverse length of pause that would affect raters' judgments has not been properly examined. This is because most previous studies have directly analyzed data from L2 learners' utterances (e.g., Derwing *et al*., 2009; Kang, 2010) in which only the already formed pausal patterns of L2 speakers' were investigated. Little consideration could be given to specific pause length parameters. Thus, it was not possible to check whether there is a threshold of pause length that might distinctively affect the raters' judgments. Second, the previous studies, by examining a relatively small amount of natural L2 utterances, have not precisely controlled other critical features in evaluation, especially segment pronunciation. Undoubtedly, segment articulation accuracy has proved to have an important influence in L2 evaluation (Cheng *et al*., 2004; Hoekje & Williams, 1992 among others). After all, those utterances with long pauses might get lower ratings, but it is not certain whether that is mainly caused by pauses or by other features like segmental pronunciation, or by their interaction.

The data in the current experiment is designed to overcome those drawbacks. In order for raters to solely depend upon pause in their evaluation, speech from a native speaker was recorded and manipulated to be used as stimuli. The key process of manipulating the stem utterances is inserting the pause of diverse length. The location of pause to be placed also varies as to whether the pause is aligned with syntactic boundary or not. The duration of the stem utterance itself is differentiated as well. Regarding rater related factors, the main criterion of recruiting was their proficiency level assuming that their age and gender are not expected to critically influence evaluation. To sum up, the four major factors are taken into consideration in the experiment: pause length, pause position, utterance length, and rater group. It should be noticed that the current setting achieves control of the influence of segmental and other prosodic features such as rhythm and intonation, which otherwise might considerably affect rating results.

Another advantage of the current study is a relatively large number of raters. To enhance statistical modelling and prediction, and discover a more general tendency for human raters, all 90 raters participated in evaluation.

The results were analyzed in two stages of statistical analysis. First, the *linear mixed-effect model* was used to check roles of four factors. Then, the *decision tree-based regression and prediction* method was adopted to clarify which of the verified factors are more influential in rating.

2. Experimental procedures

2.1. Participants

Three groups (30 EN: English native listeners, 30 KH: Korean high level listeners, and 30 KL: Korean low level listeners) of ninety evaluators who all reported that they had normal hearing participated in the experiment. Most of those in the EN group (mean age 37.5, with a range of 26 to 53) hold a master's or doctoral degree in the field of English Linguistics or TESOL, and all of those in the KH group (mean age 35.6, with a range of 25 to 52) hold a master's or doctoral degree in the field of English Linguistics or Education, and most of them majored in English related subjects in their undergraduate school, too. These two high levels are potential evaluators for real English-speaking tests.[1] Most EN and KH group participants are lecturers or professors of college level English, and the average period of teaching is 9.4 and 6.2 years, respectively. Approximately half of the KH raters have stayed in English speaking countries from 1 to 8 years, but none of them spent their critical period in the English speaking country.

The third group, KL, was composed of students (mean age 22.9, with a range of 19 to 25) whose major was not English or English education, but they were double majoring in English related areas. Most KL raters have never been abroad. Forty students originally participated as KL raters, but the four students who had stayed in English speaking countries more than two years were excluded. Then an additional three students who did not properly follow the experimental guidelines were excluded. Finally, another three, picked randomly, were excluded to make the number of raters match that of the other groups (EN and KH).

2.2. Stimuli

2.2.1. A speaker for the material

Samples from one native English speaker were selected for the test. The speaker was in his mid 30s and was born and raised in Northeast America. Recording samples were made using an AKG C520 microphone and TASCAM US-144MKII interface in the

1 Twelve of the EN, and eight of the KH raters already had official experience in evaluating English-speaking materials.

format of 16-kHz 16-bit PCM in a quiet room.

2.2.2. The test material

Two stem utterances of different length are devised as given in Table 1: The short utterance with seven syllables, and the long utterance, 15 syllables. According to his corpus data, Kendall (2013) found that the mean number of syllables per utterance was 12.04. Seven syllables is shorter and 15 is a bit longer than this mean value, meaning the pause appearance in a 15 syllable utterance would not be as unnatural as in a seven syllable utterance. However, both utterances can be produced easily without pauses. The seven-syllable length sentence with a pause, but without any complex structural units such as adjunct verb or appositive, is expected to be marked with lower scores, as it may be awkward to put a pause in such a simple short utterance.

As pausing in improper syntactic positions is a common error in low-level L2 learners' speech (Anderson-Hsieh & Venkatagiri, 1994; Pickering, 1999), two different pause positions are chosen in syntactically proper and improper positions, respectively. This is to figure out whether pause is more harshly penalized when it is misaligned with a proper syntactic boundary. Notice that the two utterances in each pair comprise the identical segmental string. Also, in both syntactically proper and improper positions, similar sequence of sounds were located immediately before and after the pause (i.e., 'a voiced sound' + /z/ + pause + /l/ + /æ/).

Table 1. Test Materials
(r : right position – pauses in syntactically proper positions,
w : wrong position – pauses in syntactically improper positions)

Utt type	Pause position	Test material	Distinction code
Long Utterance	r	Joseph the great was playing with his lanterns /PAUSE/ last Wednesday morning.	LR
	w	Joseph the great was playing with his /PAUSE/ lanterns last Wednesday morning.	LW
Short Utterance	r	John saw his lanterns /PAUSE/ last night.	SR
	w	John saw his /PAUSE/ lanterns last night.	SW

The natural pause length was ascertained by relatively recent studies of large corpus read speech data (Campione & Véronis, 2002; Clopper & Smiljanic, 2011; Kendall, 2013). The mean pause duration is 0.453 seconds for Campione & Véronis (2002), 0.455 for Clopper & Smiljanic (2011), and 0.562 for Kendall (2013). For the ease of calculations, 0.5 seconds was selected as both the shortest pause length and the interval length. Including utterances with no pause, seven different pause lengths are selected. Starting from 0.5 seconds, pauses are composed of the 0.5 second-interval except between 3 and 6 seconds (i.e., 0.5, 1, 1.5, 2, 2.5, 3, and 6 seconds). The narrower interval than 0.5 sec will generate too many stimuli tokens and too long a rating session for listeners to maintain their concentration.[2] Also, in the pilot test, there were no significant differences in raters' scores on utterances with pauses over 2 seconds, so the length of pauses was limited to 3 seconds thereafter. Lastly, according to findings by Kendall (2009) in his large-scale corpus study, the longest pause length category in English is 5 seconds. Thus, in this experiment, 6 seconds was used as the longest pause length category to find out the evaluator's reaction to extremely long pauses. The pause durations are diversified and inserted automatically through a Praat duration manipulation script in each pause position presented in Table 1.

As each test token is generated from a single stem utterance except for two types of varied pause length, all test tokens for the same length share the same articulation rate[3] (4.02 syl/sec for a short utterance and 4.36 syl/sec for a long utterance).

To summarize, a set of 30 utterances [2 types of sentence length × 15 variations (7 improper positions + 7 proper positions + 1 no-pause utterance) = 30] were generated and used for evaluation. The total number of evaluated tokens is 2,700, which were used for statistical analysis.

2.3. Durational adjustment of pre-pausal syllables

It has been widely recognized that the final syllable, or rhyme, of the word immediately preceding a pause or a major prosodic boundary is considerably lengthened (Campbell & Isard, 1991; Klatt, 1976 among others). In stimuli tokens of the current experiment, some syllables are inevitably put in that prepausal position due to artificial insertion of pause (e.g., the final syllable of word 'lanterns', in the LR sentence in Table 1). If these syllables are left without being lengthened, the unnaturalness may affect raters' evaluation confusing the target effect of pause duration. Consequently, it is attempted to factor out this prepausal syllable effect by altering their duration from the original utterance. Instead of using duration adjustment formulas frequently adopted for TTS synthesis, a simple heuristic method is applied in this experiment. First, a native English speaker was instructed to produce the target utterances without pauses (stem utterance), then with pauses as presented in Table 1. Then, as illustrated in Figure 1, the magnitude and ratio of prepausal lengthening is estimated from the utterance where a pause has originally been located in the same position. Then, the duration of the prepausal syllable has been cloned to the target utterance replacing its original duration tier. This process has been performed through a Praat duration manipulation script in a syllable-based unit. By the way, as the magnitude of temporal adjustment was restricted enough not to trigger any perceivable distortion of intonation contour, no artificial adjustment was made for intonation.

Through this adjustment, it is assumed that the unnaturalness of prepausal syllables caused by artificial pause insertion has been resolved.

2 Though raters listened to the stimuli including multiple dummy utterances, many participants mentioned that they had rated the same conditioned utterances.
3 Speaking rate was measured excluding pauses.

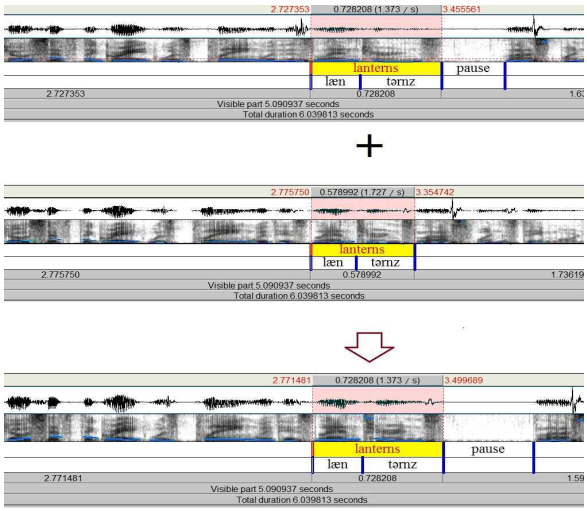

Figure 1. Procedure of prepausal syllable duration adjustment (1. The duration of the prepausal syllable from the utterance originally produced with pause is calculated. 2. The duration of the target syllable in the stimulus token is removed. 3. The prepausal duration (via procedure (1) is cloned to target stimulus (2) by re-synthesizing.)

2.4. Evaluation

Prior to each rating session, raters were given an information sheet where simple printed evaluation criteria were listed along with transcription of 30 stimuli token confounded with 42 dummy sentences. Dummy utterances were included in order to prevent too frequent appearance of consecutive tokens with pauses of similar duration as well as to alleviate monotonousness of having to hear analogous utterances.

Raters were asked to score each token's overall fluency on a scale from 0 to 9, where a rating of 0 indicated 'poorest,' and a rating of 9 indicated 'excellent.' Using a Praat MFC (Multiple Forced Choice) experiment interface, participants clicked the point scale directly on a computer. A pretest was carried out before the regular assessment, so that evaluators may get used to fluency judgment. Participants could replay the test utterances as many times as they wanted and could likewise re-rate previous utterances freely. However, most KL raters, except two who rated individually, marked on their own sheets as they rated together, in two sub-groups, listening to stimuli at the same time. Thus, if someone wanted to listen again, the other students also had to listen to the same utterance again.

Each rater evaluated a randomized list of 72 tokens. A session took approximately 13 minutes.

2.5. Analysis

Two separate but related statistical investigations were performed. First, the *linear mixed-effect models* (Laird & Ware, 1982) were fit to analyze the role of individual predictors and their interaction. Use of this method is widely expanding in various research areas thanks to its powerful functionality and wide availability of software for its application. Unnecessary assumption of variable homoscedasticity, an advantage of this method, is the main motive of employing it in the current analysis. As grouping of raters into three levels is based on their English proficiency, their individual characteristics such as age and gender could not be balanced. The mixed-effect models allow these unpredictable factors to be treated as random effects separated from the main target predictors, or fixed effects. In the current experiment, predictors that are assumed to have a systematic role in rating are classified into fixed effects (i.e., pause length, pause position and rater group), while unpredictable variables related to listeners' personal characteristics are processed as random effects (i.e., age and gender).

Once the basic functions of predictors are known, it is interesting to see which of the significant predictors plays a more crucial role in rating as compared to others. Thus, the so-called *predictor importance* is computed by adopting a statistical learning technique called *Random Forest*, an application of the decision-tree-based modeling and prediction. A benefit of this method is that it can treat variables with non-linear observations as in our rating data. The quantitative and graphical comparison of predictor importance will be given through this technique.

These two processes are implemented using the publicly available software *R* (version 3.3.2) and its packages, especially *lme4* (Bates *et al*., 2015) for the linear mixed-effect models and *caret* for the tree-based models (Kuhn, 2008).

3. Results

Table 2 is the summarized output of the linear mixed effect model fit where SCORE is the response variable varying upon main predictor variables: RATER GROUP, PAUSE LENGTH, PAUSE POSITION and UTTERANCE LENGTH as fixed effects, while INDIVIDUAL LISTENER, AGE and GENDER as random effects. Considering the random effects first, the individual listener variable is, as expected, distinctly more variable than age or gender. In other words, characteristics of each listener is in a certain degree responsible for undermining predictability of the model.

As for the fixed effects, all variables except the proficiency group are found to affect rating scores. Although, the overall result shows that the rater group alone seemed to be influential only in a marginal degree, its role is also frequently significant when interacted with other variables, as will be described in the following subsections.

Table 2. The results of linear mixed effect model

Random effects				
Groups	Name	Var	Std. Dev.	
listener	(Intercept)	1.070	1.035	
age	(Intercept)	0.003	0.059	
gender	(Intercept)	0.000	0.000	
Residual		1.217	1.103	
Number of obs: 2700, groups: listener, 90; age, 31; gender, 2				
Fixed effects				
	Estimate	Std. Err.	t	p[4]
(Intercept)	7.098	0.197	36.04	< 0.001***
levelKH	0.655	0.273	2.40	= 0.074
levelKL	0.427	0.274	1.56	= 0.194
Pau Leng	-0.494	0.012	-40.14	< 0.001***
Pau PosW	-0.816	0.043	-19.10	< 0.001***
LengthShort	-0.355	0.042	-8.36	= 0.001**

3.1. Pause length

As presented in Table 3 and Figure 2, all rater groups consistently gave lower scores as pause became longer. The correlations between mean rating scores and pause length are strong for each group (-0.861 for EN, -0.830 for KH, -0.887 for KL).

Table 3. Mean rating scores for each speaker group and correlation between mean scores and pause length

pause length (sec)	Average	EN	KH	KL
0 (no pause)	8.53	8.57	8.60	8.43
0.5	6.79	6.58	6.89	6.91
1.0	6.20	5.93	6.43	6.24
1.5	5.82	5.53	6.08	5.85
2.0	5.68	5.18	6.00	5.87
2.5	5.38	5.00	5.63	5.51
3.0	5.18	4.68	5.58	5.28
6.0	4.28	3.73	4.88	4.24
correlation	**-0.862**	**-0.861**	**-0.830**	**-0.887**

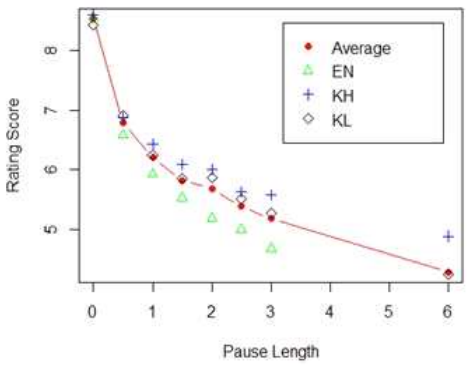

Figure 2 . Mean rating scores for each listener group on pause length

It is illustrated in Figure 2 that native speakers (EN) penalize lengthened pauses more harshly than Korean raters. All rater groups gave distinctly high scores when there was no pause, while they gave extremely low scores to utterances with an unusually long (6 sec) pause.

Although a gradual decrease in ratings due to increasing pause length has been statistically verified ($\chi^2(7)=1134.5$, $p<0.0001$), a question may arise as to how sensitively raters perceive and penalize changes of pause duration. To verify this, pairwise contrast between adjacent units of pause length was conducted and its outcome is shown in Table 4.

Table 4. Pairwise contrast on pause length based on TukeyHSD. The length '0' denotes 'no pause'. ('***' p<0.0001, '**' p<0.001, '*' p<0.01)

	0.5	1	1.5	2	2.5	3	6
0	***	***	***	***	***	***	***
0.5		***	***	***	***	***	***
1			**	***	***	***	***
1.5				NS	***	***	***
2					*	***	***
2.5						NS	***
3							***

The results indicate that most of the 0.5-second intervals influence raters significantly except for the two intervals: 1.5-2.0 sec and 2.5-3.0 sec meaning that 0.5 sec difference may not be further penalized in evaluation when pause duration is over 2 sec which is already unacceptable amount in an utterance spoken by a native speaker. However, there is no 1-second interval that does not affect the raters' scoring. More discussion on pause length is given in Section 5.

It should also be noted that the interval between no-pause (or, length '0') and the minimal pause, 0.5 sec, triggers a more distinct difference in score (i.e., 1.74 points) than any other adjacent intervals, as shown in Table 3 and Figure 2. This is clear evidence that raters are crucially influenced by pause appearance itself apart from duration of the utterance internal pause. Consequently, it could be concluded that both appearance and duration of pause significantly affect the score of spoken utterances.

3.2. Pause position

Lower rating scores were given to utterances with a pause in syntactically improper positions (mean score of 5.31) than to those in proper positions (mean score of 5.95, $\chi^2(1)= 293.44$, $p<0.0001$), and scores from each of the three rater groups as well as the group overall average showed statistically significant differences ($p<0.001$). This suggests that the position of pauses have a significant effect on evaluation in general in such a way that pauses aligned with syntactic boundaries are less penalized than misaligned pauses.

An interaction between features of pause position and rater group was found to exist ($F(2, 2694) = 17.18$, $p < 0.001$), as illustrated in Figure 3. That is, the KH group has given distinctly high scores to utterances with an aligned pause.

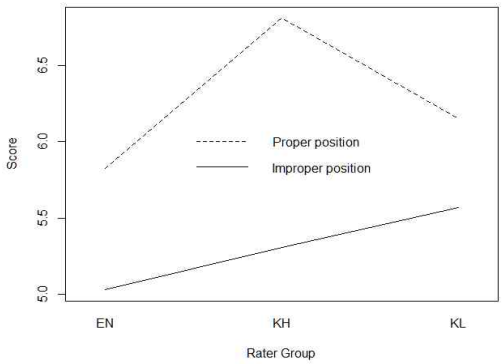

Figure 3. Interaction between Pause Position and Rater Group

3.3. Utterance length

There appeared to be higher marks given on relatively long utterances (mean score of 6) than short utterances (mean score of 5.65, $\chi^2(1)= 28.308$, $p<0.0001$).

Regarding rater groups, both of the Korean raters, KL and KH,

4 The p-values in this fitting were obtained through the Kenward-Roger method (Kenward & Roger, 1997) which approximates degrees of freedom and the *t*-distribution to get p-values. This method is available from the R package *pbkrtest*.

evaluated differently based on utterance length while no significant influence was found in the EN group's evaluation. But the Korean raters' preference on long utterances was not maintained when there is no pause within the utterance. It means that Korean raters tend to tolerate an inserted pause in relatively longer utterances.

This effect of utterance length is more distinct in KL raters than KH raters. The KL group assigned distinctly higher scores (i.e., mean 0.47 points) to long utterances regardless of whether the pause location was syntactic-boundary aligned or not, and whether its duration was long or short. It has been reported that pausing in long sentences improves listening comprehension for L2 learners (Blau, 1990; Buck, 2001). Raters of relatively lower proficiency levels like the KL group probably need more processing time for longer utterances than higher-level groups. Therefore, the KL group seemed to tolerate the pauses in long utterances regardless of direct pause-related features. However, the EN group will not need the aid of a pause to comprehend this length of sentences; hence, they give similar scores regardless of sentence length.

There was an interaction of utterance length and pause position. Listeners tend to give higher marks on long utterances in both pause positions (mean scores: long utterance - r (6.35), w (5.6), short utterance - r (6.19), w (5.02)). However, a greater gap was found for utterances with improperly positioned pauses as illustrated in Figure 4. This indicates that raters deduct fewer points on longer utterances than short utterances with the same improper pause conditions.

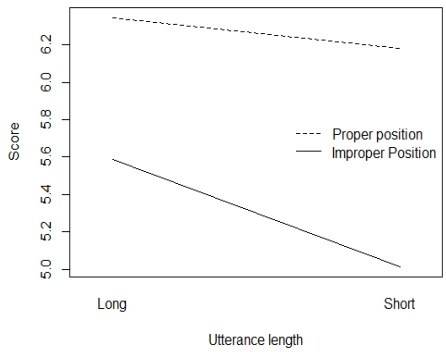

Figure 4. Interaction between utterance length and pause position

Although pause position differentiated scores of all three rater groups, when analyzed together with sentence length, the KH group showed a different rating tendency for long utterances as is shown in Table 5. They were sensitive to pause positions and consistently deducted scores for an improperly positioned pause regardless of whether the utterance is long and short. On the other hand, the EN and KL groups showed statistically less significant rating differences on long utterances. This suggests that they are more tolerant of pauses in longer sentence because the pauses in short sentences are more unnatural and abnormal in ordinary speaking situations. However, the KH group showed that they placed importance on pause position even in long utterances. This seemed to lead to statistically different results of the KH group from the other groups ($\chi^2(2)= 101.93$, $p<0.0001$).

Table 5. Interaction between pause position and utterance length for each rater group ('***' p<0.0001, '**' p<0.001, '*' p<0.01, '.' p<0.05)

utterance length, pause position	Average	EN	KH	KL
long, r-w	***	*	***	.
short, r-w	***	***	***	***

4. Regression trees and predictor importance

It has been found that features such as pause position, pause duration and utterance length significantly affect raters' evaluation. The rater group also appears to influence scores, although marginally. It will be interesting to check which of these four variables plays the more crucial role in raters' score deduction.

In the subsequent experiment, a statistical learning algorithm known as the tree-based regression was used to investigate variable importance. The reason for using this method of predictor importance instead of linear modeling is that regression trees are known to perform better than linear regression when variable relations are not assumed to be linear. It appears, as in Figure 2, that the pause related variables cannot be assumed to be linearly related with rating scores, especially when data points are concerned to each individual rating group.

In order to prevent the high variance problem that frequently deteriorate the performance of the pure tree-based regression and prediction, performance enhancing techniques known as Bagging and Random Forest are additionally adopted and their performance with the current data was compared. The procedure of these techniques can be summarized as follows:

[1] The data set is split into the training and test set, each with 1,350 data observations picked randomly.
[2] The two versions of the regression tree are fit to the training data. One with the Bagging and the other with Random Forest.
[3] Prediction performance with test data and with varying numbers of terminal nodes was monitored to find the best amount of nodes.

The step [3] above needs further elaboration: decent prediction performance is not obtained with the model trained with only one set of data set due to high variance. Thus, the training set is split into many subsets and the model parameters are calculated by averaging those subset results to lower the variance. This bagging method, also known as bootstrapping, has a problem. All of bagged trees will inevitably be highly correlated as they are from the same population. Usually highly correlated predictors are unlikely to reduce as much variance as expected when averaged. To alleviate this problem, a de-correlation method called Random Forest is adopted. Instead of using all predictors for each bagged tree, a reduced number of randomly chosen predictors are used. In the current experiment, for example, 3 out of 6 predictors are picked for building each bagged model. Trees built in this way become less homogeneous leading to a more meaningful reduction of variance.

Figure 5 shows that Random Forest (solid line) bootstrapped with less predictors outperforms the bagged model with the maximum number of predictors, confirming that de-correlation modeling is useful.

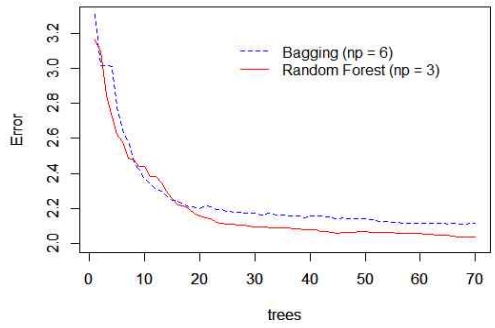

Figure 5. Performance of Random Forest. The test errors are given as a function of bootstrapped subsets of training data. The number of predictors is denoted by 'np'.

The role of each predictor in the random forest model can be represented in terms of permutation error (%incMSE in Figure 6). Its value for a predictor variable indicates the amount of MSE increase when the predictor's value is fixed to a single quantity/quality. The greater the permutation error, the more important the predictor.

Figure 6 shows that pause length is a more important predictor in evaluation than pause position or utterance length. Notice that raters' age appears to be another important predictor. But a closer examination reveals that the age is closely associated with the feature, rater group. The mean age of each group was found to be 37.5 (EN), 35.6 (KH), and 22.9 (KL), respectively. This is because KL, or Korean low-proficiency raters are mostly college students. Furthermore, the age range of raters is quite restricted and unevenly distributed. Consequently, it is not appropriate to jump to a conclusion that raters' age significantly affects their evaluation. After all, that is why we attributed such variables as age and gender, that are related to listener's biographical information, to random effects rather than fixed effects in the experiment.

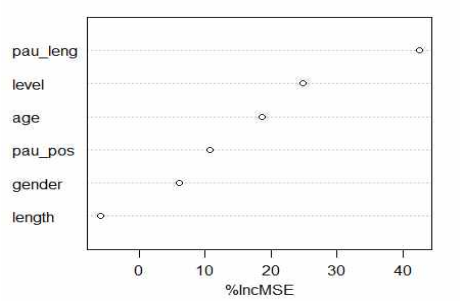

Figure 6. Predictor importance in terms of permutation error. The '%IncMSE' values denote the increase of MSE for each unused predictor.

It is interesting that the pause position and utterance length were also statistically significant predictors but their impact in rating is found to be relatively smaller than expected.

5. Summary and discussion

The aim of this study was to investigate the pause-related features affecting assessment results. All 90 listeners evaluated the controlled materials for pause length and position with two different sentence length conditions. The statistical results verified that rating is affected by such features as pause length, pause position, and utterance length. The pause appearance itself is also found to influence ratings. Although there were different characteristics by each rater group especially with respect to interaction with specific features, tendency of pause effect was shared by all three groups.

Kowal & O'Connell (2008) considered pauses over 1 second as outliers, and Kendall (2013) reported that 86 percent of pauses in his large corpus of pause data were below 1 second in length. Thus, if listeners encounter a pause exceeding an acceptable length, it can be reasonably expected they regard the pause as unsuitable as it will be deemed as abnormal and unfamiliar to encounter in an ordinary L1 speaking environment. However, the current study revealed that raters deduct scores, down to an average 7.43 points out of 9, even when the pause is shorter than 1.0 sec, it is properly positioned and the utterance is sufficiently long. It implies that an utterance with a pause of any type may not be heard as comfortably by listeners as an utterance without it.

The findings in this paper have special implications in L2 speech learning and evaluation. Since raters are influenced by pause appearance, its length and positions in evaluation, speaking test takers would rather avoid inserting pauses, proper or improper, which is likely to give an unfavorable impression to evaluators. As raters also assigned harsher scores to short utterances, especially those with a pause improperly positioned, more careful attention is needed when speakers use pauses in short utterances. In order to earn top scores on L2 speaking utterances, speakers need to try speaking with as few pauses as possible. At least, they should avoid inserting pauses longer than 1 second, particularly in improper positions, especially when the utterance is short.

There were some differences between English native evaluators and Korean potential evaluators. The EN group assigned lower scores in general, and they were more sensitively affected by pause length. The KH group enforced stricter standards on pause position than the other groups. There were also some other issues which might cause a discrepancy among rating scores. Some evaluators assigned high scores to an extremely long pause when it was properly positioned, while others assigned very low scores for those. In order to avoid inconsistent rating results from raters, a more specific rubric about pause-related issues will be needed.

It is believed that carefully designed test materials in the current experiment contributes to producing more reliable results than previous relevant research. To move forward, test materials using carefully controlled L2 speakers' utterances would provide meaningful parallels to this study. Also, a controlled experiment with more varied pause-related features like frequency of pauses or other temporal features such as speaking rate would be useful.

References

Anderson-Hsieh, J., & Venkatagiri, H. (1994). Syllable duration and pausing in the speech of Chinese ESL speakers. *TESOL Quarterly*, 28(4), 807-812.

Bates, D., Maechler, M., Bolker, B., & S. Walker (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1-48.

Blau, E. K. (1990). The effect of syntax, speed, and pauses on

listening comprehension. *TESOL Quarterly*, 24(4), 746-753.
Buck, G. (2001). *Assessing listening*. Cambridge: Cambridge University Press.
Campbell, N. W., & Isard, S. D. (1991). Segmental durations in a syllable frame. *Journal of Phonetics*, 19(1), 37-47.
Campione, E., & Véronis, J. (2002). A large-scale multilingual study of silent pause duration. *Proceedings of the 1st International Conference on Speech Prosody* (pp. 199-202). Aix-en-Provence, France.
Cheng, L., Myles, J., & Curtis, A. (2004). Targeting language support for non-native English-speaking graduate students at a Canadian university. *TESL Canada Journal*, 21(2), 50-71.
Clopper, C. G., & Smiljanic, R. (2011). Effects of gender and regional dialect on prosodic patterns in American English. *Journal of Phonetics*, 39(2), 237-245.
Derwing, T. M., Munro, M. J., Thomson, R. I., & Rossiter, M. J. (2009). The relationship between L1 fluency and L2 fluency development. *Studies in Second Language Acquisition*, 31(04), 533-557.
Hoekje, B., & Williams, J. (1992). Communicative competence and the dilemma of international teaching assistant education. *TESOL Quarterly*, 26(2), 243-269.
Kang, O. (2010). Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness. *System*, 38(2), 301-315.
Kendall, T. (2009). *Speech rate, pause, and linguistic variation: An examination through the sociolinguistic archive and analysis project.* Ph.D. Dissertation, Duke University.
Kendall, T. (2013). *Speech rate, pause and sociolinguistic variation: Studies in corpus sociophonetics*. London: Palgrave Macmillan.
Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 53, 983-997.
Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *The Journal of the Acoustical Society of America*, 59(5), 1208-1221.
Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, 32(2), 145-164.
Kowal, S., & O'Connell, D. C. (2008). *Communicating with one another: Toward a psychology of spontaneous spoken discourse*. New York: Springer.
Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5), 1-26. Retrieved from http://www.jstatsoft.org/v28/i05/ on January 2, 2017.
Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data, *Biometrics*, 38, 963-974.
Munro, M. J., Derwing, T. M., & Morton, S. L. (2006). The mutual intelligibility of L2 speech. *Studies in Second Language Acquisition*, 28(01), 111-131.
Pickering, L. (1999). *An analysis of prosodic systems in the classroom discourse of native speaker and nonnative speaker teaching assistants.* Ph.D. Dissertation, University of Florida.
Riazantseva, A. (2001). Second language proficiency and pausing a study of Russian speakers of English. *Studies in Second Language Acquisition*, 23(04), 497-526.
Rossiter, M. J. (2009). Perceptions of L2 fluency by native and non-native speakers of English. *Canadian Modern Language Review*, 65(3), 395-412.
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(01), 1-30.
Zellner, B. (1994). Pauses and the temporal structure of speech. In E. Keller (Ed.), *Fundamentals of Speech Synthesis and Speech Recognition* (pp. 41-62). Chichester: John Wiley.

- **Mi-Sun Kim**
  Department of English Linguistics
  Hankuk University of Foreign Studies
  107 Imun-ro, Dongdaemun-gu
  Seoul 02450, Korea
  Tel: 02-2173-2266
  Email: alice-ms@hanmail.net
  Fields of interest: Phonetics

- **Tae-Yeoub Jang** corresponding author
  Department of English Linguistics
  Hankuk University of Foreign Studies
  107 Imun-ro, Dongdaemun-gu
  Seoul 02450, Korea
  Tel: 02-2173-2900
  Email: tae@hufs.ac.kr
  Fields of interest: Phonetics, Speech Technology