pISSN 2005-8063
eISSN 2586-5854
2023. 3. 31.
Vol.15 No.1
pp. 47-54

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

https://doi.org/10.13064/KSSS.2023.15.1.047

Evaluation of English speaking proficiency under fixed speech rate: Focusing on utterances produced by Korean child learners of English*

Narah Choi · Tae-Yeoub Jang**

Department of English Linguistics, Hankuk University of Foreign Studies, Seoul, Korea

Abstract

This study attempted to test the hypothesis that Korean evaluators can score L2 speech appropriately, even when speech rate features are unavailable. Two perception experiments—preliminary and main—were conducted sequentially. The purpose of the preliminary experiment was to categorize English-as-a-foreign-language (EFL) speakers into two groups—advanced learners and lower-level learners—based on the proficiency scores given by five human raters. In the main experiment, a set of stimuli was prepared such that the speech rate of all data tokens was modified to have a uniform speech rate. Ten human evaluators were asked to score the stimulus tokens on a 5-point scale. These scores were statistically analyzed to determine whether there was a significant difference in utterance production between the two groups. The results of the preliminary experiment confirm that higher-proficiency learners speak faster than lower-proficiency learners. The results of the main experiment indicate that under controlled speech-rate conditions, human raters can appropriately assess learner proficiency, probably thanks to the linguistic features that the raters considered during the evaluation process.

Keywords: speech rate, speaking evaluation, L2 proficiency, fluency, prosody evaluation

1. Introduction

As suggested and verified in much research, speech rate helps distinguish between L1 and L2 speakers. Utterances produced by non-native speakers tend to be spoken more slowly than those by native speakers. Speech rate is also differentiated by language learners' proficiency levels: the lower the proficiency of a language learner, the slower their utterance (Baker-Smemoe et al., 2014; Derwing & Munro, 1997; Huang & Gráf, 2020).

Usefulness of speech rate has been found in second language (L2) studies. Arevart & Nation (1991) discovered that language learners were recognized to perform at a higher level of fluency when they practiced speaking faster during a retelling activity. However, there is no consensus on whether and how speech rate influences evaluation of L2 speakers' proficiency. For example, Flege (1988) reported that there was no significant difference in listeners' foreign accent scoring depending on speeds of Mandarin speakers' English utterances. On the other hand, Munroe & Derwing (2001), after a set of foreign-accent judgment experiments, concluded that speech rate does affect listeners' evaluation significantly. They reported that

* This work was supported by Hankuk University of Foreign Studies Research Fund of 2023.

** tae@hufs.ac.kr, Corresponding author

Received 5 March 2023; Revised 16 March 2023; Accepted 20 March 2023

raters gave the best score to speeds that are considerably faster than L2 speakers' average rate but slightly slower than native speakers' speed.

Speech rate is relatively easier to measure acoustically than other metrics; thus, its temporal characteristic has been employed as a major feature not just by human raters to perceive second language learners' fluency (De Jong et al., 2013; Kormos & Dénes, 2004) but also in quantitative assessments of second language learners' fluency using automatic speech processing technology (Cucchiarini et al., 2000; Cucchiarini et al., 2002; de Wet et al., 2009).

While it is true that speech rate is salient enough to be employed, raters, particularly untrained raters, are prone to overreliance on utterance speed in proficiency evaluations. The same may happen in automatic scoring systems whereby speed is weighted more heavily than other features. If either human raters or machines are too dependent on speech rate, proficiency test results cannot fully reflect the fundamentals of linguistic elements concerning proficiency. Undoubtedly, there are distinct linguistic domains that are important in proficiency (Saito et al., 2016); besides, complexity, accuracy, and fluency have all been emphasized for improving proficiency in language classrooms (Housen, 2009; Housen et al., 2012; Skehan, 2009).

Thus, the current study attempts to testify the hypothesis that human evaluators can score L2 speech appropriately even when the feature of speech rate is totally unavailable. Two perception experiments, preliminary and main, were conducted one after the other. The purpose of the preliminary experiment is to categorize English-as-a-foreign-language (EFL) speakers into two groups based on their proficiency level: advanced learners or lower-level learners. Then, in the main experiment, a set of stimuli were prepared in such a way that the speech rate of all data tokens were manipulated to have the same speech rate leaving all the other acoustic information untouched. Evaluators were provided with each of these stimuli tokens and requested to score on a 5-point scale. It was statistically analyzed whether the scores obtained in this way were significantly different between the two groups.

2. Experimental Data

For both experiments, data were selected from Korean-spoken English corpus (K-SEC) created by Rhee et al. (2009). Since this paper limits the scope to English proficiency differences among Korean child EFL learners, a set of 36 English sentences that were uttered by elementary school students was extracted first and further specified for each experiment as described in the following sections.

2.1. Speakers

To minimize any unwanted effects caused by dialect or gender, the speakers were narrowed down to 32 children from the original K-SEC: 16 were from Seoul (eight males, eight females) and the other 16 children were from Gyeonggi-do province (eight males, eight females). They were between 9 and 13 years old. Their mean age was 11.25 (*SD*=1.06). For the naturalness assessment, described in subsection 4.5.3, in the main experiment and further analysis, three native speakers of English aged 10–11 (two males, one female) were also included.

2.2. Sentences

According to Rhee et al. (2003), the subset of 36 English sentences from K-SEC were recorded by each speaker. The set consisted of twenty-two declarative sentences, eight interrogative sentences, five imperative sentences, and one exclamatory sentence. The number of words per sentence was 7.22 (*SD*=1.81) on average.

3. Experiment I: Proficiency Categorization

The preliminary experiment was designed to categorize the children into one of two groups based on their proficiency level: advanced learners or low-level learners. This categorization result will be utilized in the main experiment as preset proficiency data for each speaker.

The speech rate of each group was also analyzed for the purpose of finding all speakers' mean speech rate which will be used as the fixed speech rate modulation.

3.1. Data Subset

Among the data from K-SEC, 288 recordings were selected for the first experiment. These tokens were from 10 sentences out of the list spoken by 32 Korean children from Seoul and Gyeonggi-do province. A total of 10 percent of the recordings taken (i.e., 32 tokens) were missing in the original database, leaving 288 tokens. Table 1 shows a list of sentences for proficiency evaluation in this preliminary experiment.

Table 1. Sentences for proficiency evaluation in Experiment I

	No.	Sentence	Type
Part 1	1	Miss Henry drank a cup of coffee.	Declarative
	2	What are you looking for?	Interrogative
	17	It's my sister who talked to the kid.	Declarative
	19	The police took the cab to Seoul.	Declarative
	25	Did he fail the test again?	Interrogative
Part 2	3	Put your toys away right now.	Imperative
	6	Hit the ball with this bat.	Imperative
	20	What a surprise!	Exclamatory
	22	I have friends who are just like me.	Declarative
	29	You like orange juice, don't you?	Declarative1[1]

3.2. Raters

Five graduate students participated in the first experiment as the raters: three master's and two doctoral students. They were all majoring in English Linguistics at a university in Seoul, South Korea.

3.3. Methods

The experiment was conducted online using GORILLA, an online testing tool publicly available at https://gorilla.sc (Anwyl-Irvine, 2020). During the experiment, the participants were asked to evaluate each utterance token by awarding scores ranging from 1

1 Since the main part of the tag question usually consists of 'statement', this type is classified as 'declarative' instead of 'interrogative'.

(low) to 5 (advanced) while listening to the recordings. The raters were allowed to listen to each recording up to three times but could not go back and modify scores for previous utterances.

The experiment comprised two parts with a 10-minute intermission. In each session, the raters evaluated 144 utterances. To prevent learning effects, the order of recordings was pseudo-randomized in both sessions, ensuring no consecutive speakers or sentences.

3.4. Results

Using R (version 4.2.2, R Core Team, 2022), a tool for statistical analysis, *t*-tests were conducted to see the differences between low and advanced groups and correlation tests were also adopted for checking evaluation reliability.

3.4.1. Proficiency scores of each learner

After the five raters had marked the scores of 288 recordings, the mean scores of each child were calculated. Four students with the highest scores were assigned to the advanced group (adv) and five students with the lowest scores were placed in the low-level group (low). The intermediate-level learners were not considered since the advanced learners in this study were distinctly different from low-level learners, but they may not differ from intermediate learners.

Table 2 illustrates the mean scores of each speaker in the two proficiency groups. As shown, the difference between the advanced group of more than 3.5 and the lower level group of less than 2.5 is clear and statistically meaningful ($t(1,240.3)=27.98$, $p<0.001$).

Table 2. Means (and standard deviations in parentheses) of each speaker's proficiency score

Group	Speaker ID	Score
Adv	1215	4.78 (0.51)
	1112	4.06 (0.61)
	1216	3.90 (0.70)
	2211	3.80 (0.66)
	Mean	4.12 (0.73)
Low	2212	2.34 (0.70)
	2215	2.31 (0.99)
	2216	2.28 (0.64)
	2118	2.26 (0.73)
	1115	1.93 (0.80)
	Mean	2.22 (0.81)

3.4.2. Reliability: correlation between the raters

It was important to verify whether the evaluation results were reliable enough for analysis. Thus, the correlation between the raters was analyzed. As Table 3 indicates, there was a high level of correlation among the raters who scored the proficiency levels of each speaker.

Table 3. Correlation between the raters

	S2	S3	S4	S5
S1	0.908	0.891	0.938	0.935
S2		0.879	0.925	0.880
S3			0.876	0.931
S4				0.925

All pairwise cases of correlation are statistically significant ($p<0.01$).

3.4.3. Analysis of speech rate

Speech rate, or articulation rate, was measured as the number of syllables per second (syl/sec), excluding the duration of internal pauses. Table 4 provides more detailed information on the speech rate of each speaker.

Table 4. Means (and standard deviations in parentheses) of speech rate for each speaker (syl/sec)

Group	Speaker ID	Speech rate
Adv	1112	3.63 (0.48)
	1215	4.01 (0.35)
	1216	3.18 (0.30)
	2211	3.14 (0.36)
	Mean	3.49 (0.51)
Low	1115	2.35 (0.36)
	2118	2.87 (0.39)
	2212	2.89 (0.36)
	2215	2.45 (0.34)
	2216	2.28 (0.44)
	Mean	2.57 (0.46)

As Figure 1 indicates, advanced learners spoke faster than low-level learners ($t(248)=13.88$, $p<0.001$), confirming that more fluent speakers speak faster in general. Meanwhile, the mean speech rate of all the speakers was 3.0305 syl/sec, which will be used to fix the speech rate of all stimuli tokens, in the main experiment.

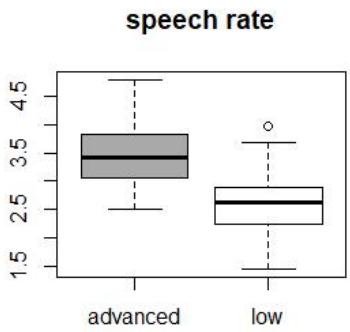

Figure 1. Boxplot of speech rate per group.

4. Experiment II: Main Experiment

Based on the results of the preliminary experiment, nine Korean EFL learners were selected: four children with the highest scores were placed in the advanced group and five children with the lowest scores were assigned to the low-level group.

The main experiment then was conducted to investigate whether the raters could perceive proficiency differences between the two groups when the speech rate of all the recordings was modulated to reflect the same speed. To ensure the reliability of evaluations, naturalness assessment followed the test to see whether the raters could distinguish original speech sounds from artificially modulated ones.

Also, a post-questionnaire was designed to determine the linguistic elements that the raters mainly considered during their proficiency evaluations, the procedure and result of which will be discussed in Section 5.

4.1. Stimuli

For rating proficiency, 250 utterances were examined. A total of

36 English utterances spoken by four advanced learners and five low-level learners were employed. There should have been 324 tokens, but 12 tokens of the advanced group and 62 of the low-level group were missing in the original database, resulting in 132 tokens of the advanced group and 118 tokens of the low-level group, respectively.

In order to remove the speech rate effect in evaluation, each original sound token was modified to reflect the same speech rate of 3.0305 syllable/second, which was the mean of all speakers. For each token, the whole utterance duration was accordingly stretched or compressed while keeping other factors such as pitch and intensity intact. The length of the internal pause, if exists, was also increased or decreased at a constant rate. This procedure was completed using Praat Vocal Toolkit (Corretge, 2012-2022).

4.2. Participants

The same five graduate students participated in the main experiment. To prevent them from remembering the characteristics of certain speakers or sentences, the this experiment was conducted approximately four weeks after the preliminary experiment.

Additionally, a group of five professionals in English education also participated in the main experiment as the evaluators: two middle school English teachers and three high school English teachers. They had all taught English in Seoul for at least three years. The reason for adding this group of evaluators is to check if professionals in the actual English education field will apply different criteria in evaluation of L2 spoken English.

4.3. Evaluation

The main experiment was also conducted using the online tool named GORILLA available at https://gorilla.sc. The experiment consisted of two sessions with a 10-minute break between each. To prevent learning effects, the order of recordings was pseudo-randomized.

During the first session, the raters scored proficiency for 250 utterances. The same proficiency evaluation methods used in the previous experiment was applied. The raters awarded scores ranging from 1 (low) to 5 (advanced) for each child's utterance.

Participants then moved onto naturalness assessment, where they identified whether each utterance had been modified or not. An utterance that they thought was natural or original sound was marked as 1; an utterance that they thought was artificially adjusted was marked as 2.

4.4. Segmentation

To measure speech rate and analyze segmental errors and other related metrics, speech segmentation was conducted using Montreal Forced Aligner (MFA) (McAuliffe et al., 2017), which automatically segments utterances and provides word and phone boundaries. Then, segmentation was checked and calibrated manually via Praat (Boersma & Weenink, 2021), a widely-known speech analysis tool.

As Figure 2 shows, five tiers were created, each tier indicating the following information: words, phones, segmental errors, and vocalic and consonantal intervals, respectively. Errors were annotated in the third tier mostly following the annotating convention suggested in L2-ARCTIC corpus (Zhao et al., 2018). Vocalic intervals were measured from the onset of the vowel to the offset of the vowel; consonantal intervals were measured from the onset of the consonant to the offset of the consonant. Glide was treated as a consonant in the onset position and was treated as a vowel in the other positions. The speech rate of each sound was calculated using the fourth tier. For acoustic analysis, the fifth tier was used where intervals of consecutive vowels or consonants merged into one interval.

The annotation illustrates the following information: target phone labels, perceived (i.e., actually pronounced) phone labels, error types, words. Phones were transcribed using CMUbet for the convenience of easy keyboard input.

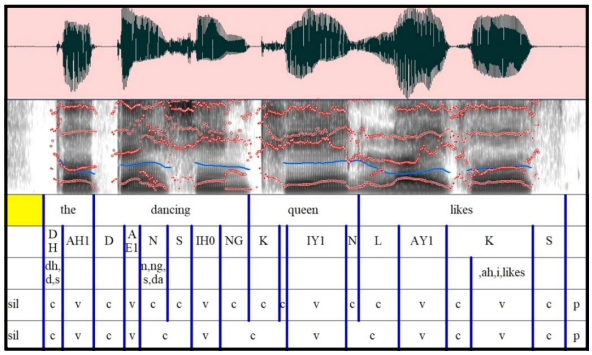

Figure 2. Example of speech segmentation in English.

4.5. Results

Above all, a linear regression modeling was performed to check whether there is any difference between scores of two proficiency groups and between two experiments. The response variable 'SCORE' was modeled by predictors 'Proficiency-Group' and 'Experiment' through the widely used formula available at R (R Core Team, 2022): *lm* (*SCORE~Proficiency-Group+Experiment*). As a result, the proficiency difference between advanced level and lower level was confirmed (t=19.88, p<0.001) while no difference was found between two experiments (t=47.93, p=0.377). More detailed results and their interpretation are as follows.

4.5.1. Proficiency scores of each learner

Both graduate students and teachers rated the proficiency of 250 utterance tokens. Table 5 illustrates the means (and standard deviations) of each speaker's proficiency score.

The differences between the advanced and low-level groups were statistically meaningful [students: t(1,235.9)=31.16, p<0.001, teachers: t(1,240.3)=27.98, p<0.001]. This result confirmed the hypothesis that human raters can accurately evaluate proficiency under controlled speech rate conditions.

Table 5. Means (and standard deviations in parentheses) of each speaker's proficiency score

Group	Speaker	Score (by students)	Score (by teachers)
Adv	1215	4.45 (0.65)	4.35 (0.78)
	1112	3.51 (0.66)	3.84 (0.86)
	1216	3.52 (0.66)	3.84 (0.84)
	2211	3.33 (0.7)	3.6 (0.96)
	Mean	3.68 (0.83)	3.89 (0.91)
Low	2212	2.3 (0.73)	2.64 (0.79)
	2215	2.66 (0.8)	2.75 (0.83)
	2216	2.34 (0.72)	2.51 (0.82)
	2118	1.85 (0.85)	2.08 (0.88)
	1115	2.03 (0.76)	2.35 (0.92)
	Mean	2.23 (0.81)	2.48 (0.88)

There were some minor changes in the means and standard deviations of proficiency scores in the second experiment. Compared to the first experiment, the mean proficiency scores of advanced learners in the second experiment tended to be slightly lower. The scores of low-level learners remained relatively constant.

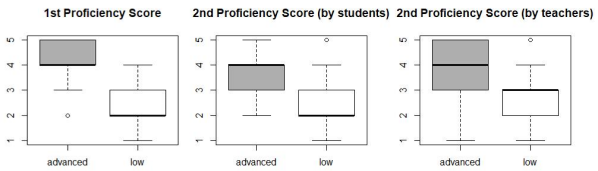

Figure 3. Boxplots of proficiency score per group in each experiment.

As Figure 3 illustrates, this tendency was partly because graduate school students were more reluctant to give a score of 5 to advanced learners compared to the preliminary experiment. A wider range of scores may affect the lowered mean score of the advanced group by the teachers. It should also be noted that one of the teachers gave the lowest score (i.e., 1) to the five utterances of speaker 2211. These five scores may be considered outliers as the teacher graded speaker 2211 a score of 3.4 on average. Removing these five scores adjusted the mean score to 3.8.

4.5.2. Reliability: correlation between the raters

The correlation between the raters was analyzed in the second experiment to test reliability. There was a strong correlation between the raters, ranging from 0.811 to 0.986. Tables 6 and 7 show the correlation coefficients between graduate school students and between teachers, respectively. Thus, it can be concluded that all the human raters were able to appropriately differentiate the two proficiency groups without having to depend on the feature of speech rate.

Table 6. Correlation between the raters (graduate students)

	S2	S3	S4	S5
S1	0.977	0.945	0.980	0.970
S2		0.937	0.972	0.959
S3			0.965	0.986
S4				0.980

All pairwise cases of correlation are statistically significant (p<0.01).

Table 7. Correlation between the raters (teachers)

	T2	T3	T4	T5
T1	0.972	0.845	0.926	0.981
T2		0.837	0.927	0.976
T3			0.962	0.811
T4				0.898

All pairwise cases of correlation are statistically significant (p<0.01).

Correlation between scores of the two experiments, preliminary and main, was also measured to verify whether participants rated proficiency scores in the same pattern, regarding or regardless of the speed of the utterances. As Table 8 indicates, correlation coefficients were consistently over 0.8, indicating the reliability of the experiments.

Table 8. Correlation between the experiments

	Exp II (S)	Exp II (T)
Exp I	0.875	0.819
Exp II (S)		0.879

All pairwise cases of correlation are statistically significant (p<0.01).
S, graduate students; T, teachers.

4.5.3. Naturalness analysis

It should be noted that, in real human speech, a change in speaking speed does not imply that duration of segmental components such as each consonant and vowel increase or decrease at the same rate. For example, Pickett (1999:147) states that vowel absorbs more temporal change as consonant movements usually attain a specific occlusion or narrow constriction. Consequently, a question may be raised as to whether the raters' perception might be influenced by unnaturalness of sitmuli in which speech rate was modulated by increasing or decreasing the length of segments at a consistent rate regardless of their types or contexts. In other words, children did not re-record utterances at the desired speed; rather, each utterance token was acoustically manipulated to reflect the same speech rate.

Therefore, to enhance the reliability of the results of the main experiment, it was important to verify whether the raters noticed differences between original tokens and modulated ones. If the raters had noticed that modulated utterances were unnatural, there might have been unexpected effects on their proficiency evaluations.

A total of 88 recordings were selected to test naturalness: 44 tokens were natural sounds and the other 44 were modified sounds. For the natural sounds, 12 children (nine Korean learners and three native speakers) spoke four sentences. Tokens of native speakers were included since the differences between original and modulated sounds might be easier to detect compared to those of language learners. Four declarative sentences in Table 2 except No. 22 were employed. Some data were missing: four utterances of low-level learners and another four of native speakers were missing among 96 recordings, resulting in 88 remaining tokens. For each token, participants were forced to click a button: either 'natural' or 'unnatural'.

Tables 9 and 10 summarize the result, illustrating the accuracy, hit rates, and false alarm rates of the participants. In these tables, 'modified' means speech-rate manipulated tokens while 'original' refers tokens without such modification. Thus, when a listener heard an 'original' token and his/her response was 'natural', this case was counted as a 'Hit'; if the response was 'unnatural', it is counted as an 'Incorrect'. Likewise, when a listener was given a 'modified' token and his/her response was 'natural', this case was counted as a

'False Alarm'; if the response was 'unnatural', it is counted as a 'Correct'.

The accuracy of both graduate students and teachers was approximately 50% or random in other words, meaning it was difficult for raters to distinguish original sounds from modified tokens. A high level of false alarm rates in both groups indicated that many modulated recordings sounded natural or they were not distinguishable from natural, i.e., unmodified tokens. In brief, it can be safely inferred that the process of artificially shrinking or stretching the duration of each speech token to unify the speech rate did not affect the listener's evaluation.

Table 9. Accuracy, hit rates, and false alarm rates of graduate students' responses

	S1	S2	S3	S4	S5
Accuracy	44%	51%	53%	49%	51%
Hit	33	37	30	33	40
False alarm	38	36	27	34	39
Incorrect	11	7	14	11	4
Correct	6	8	17	10	5
Hit rate (H)	0.85	0.82	0.64	0.77	0.89
False alarm rate (FA)	0.86	0.82	0.61	0.77	0.89

Hit: response–natural stimuli–original
False alarm: response–natural stimuli–modified
Incorrect: response–unnatural stimuli–original
Correct: response–unnatural stimuli–modified
H: Hits/(Hits+Incorrects)
FA: False Alarms/(False Alarms+Corrects)

Table 10. Accuracy, hit rates, and false alarm rates of teachers' responses

	T1	T2	T3	T4	T5
Accuracy	49%	49%	50%	48%	43%
Hit	32	32	29	30	31
False alarm	33	33	29	32	37
Incorrect	12	12	15	14	13
Correct	11	11	15	12	7
Hit rate (H)	0.73	0.73	0.66	0.68	0.7
False alarm rate (FA)	0.75	0.75	0.66	0.73	0.84

Hit: response–natural stimuli–original
False alarm: response–natural stimuli–modified
Incorrect: response–unnatural stimuli–original
Correct: response–unnatural stimuli–modified
H: Hits/(Hits+Incorrects)
FA: False Alarms/(False Alarms+Corrects)

5. Discussion

Based on the results of Experiment 2, it can be inferred that human raters do not rely unduly on speech rate when they evaluate L2 speech. A subsequent question is: what other characteristics do they base their evaluation on? To obtain an approximate answer to this question, a brief survey was conducted with the 10 evaluators (5 graduate students and 5 teachers) right after the second experiment. The questionnaire was divided into two fields, accuracy and fluency, and asked what basis the evaluation was mainly conducted in each field. Each field contains 6 linguistic items that can possibly be used as evaluation criteria. These items were picked by authors of this paper based on various previous research on L2 spoken language evaluation. The participants could choose multiple items for each question. They rated the sounds without transcription while evaluating English proficiency. Tables 11 and 12 describe the survey results on accuracy and fluency, respectively, with the top three items most selected by both rater groups highlighted in gray.

Table 11. Survey results on linguistic elements of accuracy in evaluating English proficiency

Items	S	T	Total
1. When English consonants or vowels are replaced with Korean segments (i.e., negative transfer occurs).	5	3	8
2. When consonants or vowels are uttered incorrectly but there is no problem understanding the meaning (substitution, insertion, deletion, etc.).	4	4	8
3. When consonants or vowels are uttered incorrectly to the extent that it interferes with understanding the meaning.	5	5	10
4. When consonants or vowels are lengthened.	3	1	4
5. When grammatical errors such as subject-verb agreement occur (e.g., singular/plural disagreement).	1	0	1
6. When a word itself is omitted or uttered out of context.	1	3	4

S, graduate student group; T, teacher group.

Table 12. Survey results on linguistic elements of fluency in evaluating English proficiency

Items	S	T	Total
1. When the stress pattern of a word is irregular and inconsistent.	3	4	7
2. When the intonation of a sentence is too monotonous and/or of no regular pattern.	4	5	9
3. When a sentence is uttered without proper pauses.	1	1	2
4. When too many pauses are inserted.	5	5	10
5. When hesitation or meaningless fillers frequently appear.	3	3	6
6. When a sentence is uttered either too slowly or too quickly.	3	3	6

S, graduate student group; T, teacher group.

Regarding accuracy, both graduate students and teachers considered segmental errors (including negative transfer) as one of the most important factors when deciding on accuracy, regardless of whether meaning loss occurred or not. While graduate students weighed segmental lengthening (Item 4) more greatly than teachers, teachers were more concerned about the insertion or deletion (Item 6) of words out of context. Overall, it appears that the two evaluator groups are applying similar criteria to determine accuracy in English pronunciation.

Human raters appear to have noted whether utterances contained segmental errors. When our L2 speech data were analyzed more elaborately, the number of segmental errors committed was significantly higher for low-level students as shown in Table 13, which indicates that the number and ratio of segmental error types in each.

Table 13. The number and ratio of segmental error types per proficiency group

	Adv. (399, %)	Low (577, %)
Substitution	308 (77)	423 (73)
Insertion	38 (10)	88 (15)
Deletion	49 (12)	61 (11)
Assimilation	4 (1)	5 (1)

A notable difference between the two groups was the number of insertion errors. Advanced learners made 0.29 (*SD*=0.47) insertion errors per utterance, whereas low-level learners made 0.74 (*SD*= 0.89) insertion errors per utterance. This difference between groups was highly significant ($t(173.12)=-4.09$, $p<0.001$). This implies that low-level EFL child learners are more likely to insert a vowel after a single consonant or between consonant clusters as a means of syllabification.

The results are generally in line with the findings of previous studies (Saito et al., 2016; Trofimovich & Baker, 2006; Yang & Chung, 2017). Less proficient learners also more frequently inserted vowels compared to more proficient learners. These two values reflected the differences in proficiency between the groups.

There was consensus among responses regarding fluency as shown in Table 12. In both rater groups, stress pattern (Item 1), intonation (Item 2), and the number of pauses (Item 4) were ranked high. It is interesting that improper pauses (Item 3) has not been picked by either group possibly due to relatively short sentence length of child speech. Item 5 had a relatively lower, if not lowest, number of responses as the utterances were controlled speech; as such, fewer fillers were inserted. Another relatively smaller number of raters chose Item 6 apparently because the speech rate had been adjusted to reflect the same speed in the second experiment.

Fluency had more complicated aspects. Human raters noted word stress and intonation patterns and the number of inserted pauses when evaluating proficiency, as also found by Kang (2010). This study suggested tentative acoustic measures reflecting this fluency domain. Intensity differences between stressed and unstressed vowels were significant in both the advanced and low-level groups, meaning that this measure cannot indicate proficiency. Differences in the duration of stressed and unstressed vowels were indicators of proficiency in this study. This study also suggests pitch contour lines regarding intonational characteristics such as declination and boundary tones contribute to proficiency judgment. The advanced group could be distinguished from low-level group through the number of inserted pauses, as many other studies have suggested (Bosker et al., 2013; Kang et al., 2010; Kim, 2017).

In brief, it is inferred that, without any clue of speech rate, human raters were able to appropriately evaluate child EFL learner proficiency based on various major linguistic features in both fileds of accuracy and fluency. Investigating the detailed role and weight of each characteristic is certainly beyond the scope of the current study and should be explored in further research.

6. Summary and Conclusion

The current study investigates whether human raters appropriately evaluates Korean children's English proficiency when listening to sounds where the speech rate was adjusted and kept constant to maintain the uniform speed. To test the hypothesis, two experiments were designed and conducted focusing on English utterances spoken by Korean child EFL learners.

The results of the first experiment show that children of lower proficiency speak slower than advanced learners. These results are in line with existing studies on the relationship between proficiency and speech rate (De Jong et al., 2013; Derwing & Munro, 1997; Huang & Gráf, 2020; Jang, 2009; Kormos & Dénes, 2004).

The hypothesis proposed in this paper was confirmed by the results of the main experiment, which indicates that under controlled speech rate conditions, human raters can accurately assess learner proficiency, thanks to linguistic features that the raters considered during the evaluation process.

There are limitations to this study that should be addressed in future studies. Firstly, the amount of data used was relatively small. In the second experiment, 250 tokens spoken by nine children (four advanced learners and five low-level learners) were analyzed. This small number of speakers may not be enough to compensate for individual differences.

Secondly, as only Korean evaluators participated in experiments. It is premature to generalize the current results. Further assessment verification with wide range of evaluators including native speakers is desired.

Additionally, the post-experimental questionnaire may not have properly reflected the raters' actual proficiency evaluation criteria. The evaluators holistically assessed Korean EFL learners' proficiency in this study. If the same raters participated in an experiment where they analytically awarded proficiency scores based on individual criteria, a discrepancy may be found between the results of the holistic and analytic assessments.

Despite these limitations, this paper has two implications: from a pedagogical perspective, a variety of linguistic elements in addition to speech rate should be emphasized both in assessment and in language classrooms. EFL learners can improve their proficiency when they are taught to focus on both accuracy and fluency instead of simply trying to speak faster. Regarding automatic speech evaluation, developers must take care not to rely too heavily on speech rate. More reliable results can be obtained if acoustic variables in automatic scoring reflect raters' evaluation criteria by reassigning the weights of variables.

References

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388-407.

Arevart, S., & Nation, P. (1991). Fluency improvement in a second language. *RELC Journal*, *22*(1), 84-94.

Baker-Smemoe, W., Dewey, D. P., Bown, J., & Martinsen, R. A. (2014). Does measuring L2 utterance fluency equal measuring overall L2 proficiency?: Evidence from five languages. *Foreign Language Annals*, *47*(4), 707-728.

Boersma, P., & Weenink, D. (2021). Praat: Doing phonetics by computer (version 6.2.01) [Computer program]. Retrieved from http://www.praat.org/

Bosker, H. R., Pinget, A. F., Quené, H., Sanders, T., & de Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing*, *30*(2), 159-175.

Corretge, R. (2012-2022). Praat vocal toolkit [Computer software]. Retrieved from https://www.praatvocaltoolkit.com

Cucchiarini, C., Strik, H., & Boves, L. (2000). Quantitative assessment

of second language learners' fluency by means of automatic speech recognition technology. *The Journal of the Acoustical Society of America*, *107*(2), 989-999.

Cucchiarini, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *The Journal of the Acoustical Society of America*, *111*(6), 2862-2873.

De Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, J. H. (2013). Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics*, *34*(5), 893-916.

de Wet, F., Van der Walt, C., & Niesler, T. R. (2009). Automatic assessment of oral language proficiency and listening comprehension. *Speech Communication*, *51*(10), 864-874.

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility. *Studies in Second Language Acquisition*, *19*(1), 1-16.

Flege, J. E. (1988). Factors affecting degree of perceived foreign accent in English sentences. *The Journal of the Acoustical Society of America*, *84*(1), 70-79.

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, *30*(4), 461-473.

Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy and fluency. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 1-20). John Benjamins. Amsterdam, the Netherlands.

Huang, L. F., & Gráf, T. (2020). Speech rate and pausing in English: Comparing learners at different levels of proficiency with native speakers. *Taiwan Journal of TESOL*, *17*(1), 57-86.

Jang, T. Y. (2009, November). Automatic assessment of non-native prosody using rhythm metrics: Focusing on Korean speakers' English pronunciation. *Proceedings of the 2nd International Conference on East Asian Linguistics (ICEAL 2)*. Vancouver, BC.

Kang, O. (2010). Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness. *System*, *38*(2), 301-315.

Kang, O., Rubin, D., & Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English. *The Modern Language Journal*, *94*(4), 554-566.

Kim, M. S. (2017). *The effects of pause and speech rate in evaluating English speech* (Unpublished Doctoral dissertation). Hankuk University of Foreign Studies, Seoul, Korea.

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, *32*(2), 145-164.

Munro, M. J., & Derwing, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech: The role of speaking rate. *Studies in Second Language Acquisition*, *23*(4), 451-468.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017, August). Montreal forced aligner: Trainable text-speech alignment using kaldi. *Proceedings of Interspeech 2017* (pp. 498-502). Stockholm, Sweden.

Pickett, J. M. (1999). *The acoustics of speech communication: Fundamentals, speech perception theory, and technology*. Boston, MA: Allyn and Bacon.

R Core Team. (2022). R: A language and environment for statistical computing (version 4.2.2) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org

Rhee, S. C., Lee, S. H., Kang, S. K., & Lee, Y. J. (2003). Design and construction of Korean-spoken English corpus (K-SEC). *Malsori*, *46*, 159-174.

Saito, K., Trofimovich, P., & Isaacs, T. (2016). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics*, *37*(2), 217-240.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30*(4), 510-532.

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.

Yang, S. H., & Chung, M. (2017, June). Linguistic factors affecting evaluation of L2 Korean speech proficiency. *Proceedings of the 7th ISCA Workshop on Speech and Language Technology in Education (SLaTE 2017)* (pp. 53-58). Stockholm, Sweden.

Zhao, G., Sonsaat, S., Silpachai, A., Lucic, I., Chukharev-Hudilainen, E., Levis, J., & Gutierrez-Osuna, R. (2018, September). L2-ARCTIC: A non-native English speech corpus. *Proceedings of Interspeech 2018* (pp. 2783-2787). Hyderabad, India.

• **Narah Choi**
Graduate (MA) Student, Dept. of English Linguistics
Hankuk University of Foreign Studies
107 Imun-ro, Dongdaemun-gu, Seoul 02450, Korea
Tel: +82-2-2173-3119
Email: nrchoi993@hufs.ac.kr
Fields of interest: Experimental phonetics, EFL education

• **Tae-Yeoub Jang,** Corresponding author
Professor, Dept. of English Linguistics
Hankuk University of Foreign Studies
107 Imun-ro, Dongdaemun-gu, Seoul 02450, Korea
Tel: +82-2-2173-3119
Email: tae@hufs.ac.kr
Fields of interest: Acoustic phonetics, Speech processing