7th ISCA Workshop on Speech and Language Technology in Education
25-26 August 2017, Stockholm, Sweden

Investigation of teacher-selected sentences and machine-suggested sentences in terms of correlation between human ratings and GOP-based machine scores

Yutaka Yamauchi[1], *Junwei Yue*[2], *Kayoko Ito*[3], *Nobuaki Minematsu*[2]

[1]Tokyo International University, Japan
[2]The University of Tokyo, Japan, [3]Kyoto University, Japan

yyama@tiu.ac.jp, {Jwyue, mine}@gavo.t.u-tokyo.ac.jp, ito.kayoko.6m@kyoto-u.ac.jp

Abstract

This study investigated relationships between teacher-selected stimulus sentences and machine-suggested ones in terms of the correlation between human ratings and GOP-based machine scores. In assessing shadowed speech consisting of 55 sentences recorded by 125 Japanese learners of English, it was examined which sentence combinations out of the 55 sentences could maximize the correlation between automatic scores and human ratings. A veteran teacher selected 10 sentences based on criteria such as sentence length, grammar, pronunciation, and prosodic features. The shadowed speech of the 10 sentences were manually rated by two native speakers of English focusing on pronunciation, prosody and lexical access (whether shadowing is done adequately or not after each word or phrase is identified). The same shadowed utterances were automatically assessed by the DNN-based GOP procedure. A significantly high correlation (r=0.738, p<.01) was found between manual ratings and automatic scores. Then groups of sentences were selected out of the original 55 sentences by greedy search (full search) so that correlation between their DNN-GOP scores and manual ratings of the selected 10 sentences could be maximized, and the top-ranked five combinations of groups of sentences were listed up. The number of shared sentences between the 10 teacher-selected sentences and machine-suggested ones in the top-ranked five combinations was only one, which was much smaller than expected, and thus the teacher's strategies in selecting stimulus sentences turned out to be inappropriate in terms of maximizing correlation. Examining the sentence sets suggested by the machine revealed that some particular sentences frequently appeared in the sets and seemed to have something to do with maximizing correlation. Hence, the present study compared the 10 sentences selected by the teacher and his selection criteria with the sentence sets suggested by the machine and discussed what kind of sentence should be chosen to improve reliability in automatic assessment.

Index Terms: shadowing, automatic assessment, DNN-based GOP, teacher-selected stimulus sentence, machine-suggested stimulus sentence

1. Introduction

Shadowing is a simultaneous oral reproduction task. In shadowing learners are requested to listen to and comprehend model utterances usually read by a native speaker and simultaneously reproduce them orally as quickly and accurately as possible [1]. Learners have to quickly reproduce the utterances. When they try to shadow, the native speaker's sound images still remain in the learners' auditory memory. Therefore, they can easily imitate pronunciation, rhythm and intonation and then gradually get accustomed to speaking the target language with accurate pronunciation and prosodic features like native speakers.

Shadowing practice is expected to not only enhance listening comprehension and speaking skills, but also promote learners' language processing to become more accelerated and automatized. As the model utterances are spoken at a faster rate, learners have to speed up in decoding auditory input information, comprehending the message and orally reproducing what they have heard. As a result, their language processing is thought to be changed from controlled into automatized. Shadowing is thought to include complex perception-production interaction and automatic semantic and syntactic processing [2-4]. Recent studies have reported the effectiveness of shadowing practice in second language (L2) learning in terms of listening comprehension skills [5-6], pronunciation, intonation and fluency [7-8] and overall proficiency [9].

2. Assessing shadowed utterances

Although shadowing practice seems to be prominent in developing L2 aural and oral skills, a serious problem lies in how to assess this performance objectively. In many cases the evaluator has to listen to the recorded shadowed utterances repeatedly while checking the script and calculate the ratio of the number of syllables or words correctly reproduced to the total number of syllables or words in the target passage [10]. This procedure is too time- and energy-consuming for teachers to implement this task in daily classroom activities.

To reduce the rater's burden, an automatic evaluation system was developed by our research group using HMM-based phoneme posteriors called GOP (Goodness of Pronunciation) [11]. In this system a computer can automatically compare learners' shadowed utterances with sounds expected from word strings in model utterances using an acoustic model stored in the PC at a phoneme level. This system can analyze and evaluate shadowed utterances and give scored feedback to the learner [12]. A significantly high correlation (r=0.82, p<.01) between GOP scores obtained from this system and manual scores by veteran language instructors was found. The GOP scores were also observed to be highly correlated with overall proficiency scores (r=0.84, p<.01) measured by TOEIC (Test of English for International Communication), one of the most popular standardized proficiency tests by ETS (Educational Testing Service). Thus the validity of the automatic evaluation system was confirmed [13].

10.21437/SLaTE.2017-6

3. Stimulus sentence selection

Selecting a limited number of proper sentences as stimulus sentences out of original passages is crucial to save time and energy in evaluating shadowed utterances for manual scoring. It also makes automatic assessment more effective. This study aims to investigate relationships between teacher-selected stimulus sentences and machine-suggested ones in terms of the correlation between human ratings and GOP-based machine scores. In other words, it is examined (1) which sentence combinations out of original sentences can maximize correlation between automatic scores and human ratings, (2) how teacher-selected sentences are different from ones suggested by the machine and (3) what kind of characteristics are observed in sentences forming combinations for maximizing correlation.

4. Speech data collection

Shadowed speech consisting of 55 sentences recorded by 125 Japanese learners of English were collected. An online recording site was developed for this data collection. The participants were requested to shadow the 55 model utterances without viewing any manuscripts or transcripts. Each model utterance was shadowed four times, and the fourth recorded utterance was used for data analysis. Four-time repetition was thought to get the learners accustomed to this task and lead to their best performances. Prior to shadowing and recording, they were asked to view an instruction page and practice shadowing and recording on the web.

5. Teacher-selected sentences

A veteran instructor teaching English over 30 years selected 10 sentences out of the original 55 sentences based on criteria such as sentence length, grammar, pronunciation and prosodic features as in Table 1. As sentence length becomes longer and more grammatical expressions like an embedded relative clause are included in stimulus sentences, syntactic and semantic difficulty increases in sentence processing. A sentence with a tag question or an alternative question sentence both require learners to choose proper intonation out of rising and falling options. As a result such complex sentences are expected to effectively differentiate good shadowers from those who are not. After the ten sentences were listed up, another veteran instructor checked these sentences based on her teaching experience and agreed that these sentences were appropriately selected as stimuli to differentiate learners' performances.

The shadowed utterances of the 10 sentences were manually rated by two native speakers of English focusing on three criteria: pronunciation, prosody and lexical access (whether shadowing is done adequately or not after each word or phrase is identified). The manual rating was conducted with the five-point Likert scale ranging from 1 (worst) to 5 (best); a full mark was 15, and the worst score was 3 in total.

The same shadowed utterances were automatically assessed by the DNN (Deep Neural Network)-based GOP procedure. The reason for employing the DNN-based GOP in the present study is that DNN-based acoustic models have been reported to have better accuracy than traditional acoustic models, as long as a very large amount of data is provided for machine training [14-15].

A significantly high correlation (r=0.738, p<.01) was found between manual ratings and automatic scores of the 10 sentences selected by the teacher.

6. Machine-suggested sentences

To select N sentences that can maximize the correlation between human ratings and machine scores, all the possible combinations of N sentences out of the 55 ones should be examined, and then the best combination can be detected. However, if we use ten for N, the number of possible combinations is almost infinite. Although it may not be theoretically impossible, implementing greedy search (full search) on such a large scale seems to be fairly impractical. For this reason we took three for N. However, we did not use only the top one combination but used the top 5, 10, 100 and 1000 combinations for further analysis in this investigation as shown in Tables 2 and 4.

Table 1: *Mean scores of attributes of 10 teacher-selected sentences*

	Ten teacher-selected sentences	sentence length		sentence complexity	T-unit length		prosody	automatic score
		word number	syllable number	T-unit	word number	syllable number	liaison	DNN-GOP
1	I'M STUDYING PHOTOGRAPHY,TOO. SHALL WE EXCHANGE SOME RECENT PHOTOS WE'VE TAKEN AND DISCUSS THEM ON THE INTERNET?	18	26	2	9	13	0	0.318
2	THE BOY SAID THAT IT HAD ALREADY BEEN BROKEN BEFORE HE AND HIS FRIEND WENT TO THE HOUSE.	18	20	1	18	20	2	0.299
3	IT WAS YOU WHO KICKED THE DOOR OPEN, WASN'T IT?	10	10	1	10	10	1	0.274
4	DID YOU JUST WANT TO HAVE A BIT OF FUN, OR WERE YOU TRYING TO GET SOME MONEY?	8	19	1	8	19	3	0.259
5	THEN ON FEBRUARY FOURTEENTH, TWO HUNDRED SEVENTY AD, A MAN NAMED VALENTINE WAS KILLED BY THE ROMANS BECAUSE OF HIS CHRISTIAN BELIEFS.	12	37	1	12	37	1	0.271
6	ALL OF A SUDDEN VALENTINE'S DAY BECAME A BIG HOLIDAY FOR PEOPLE WHO MADE AND SOLD CARDS.	17	24	1	17	24	2	0.234
7	THE HOSPITAL DOCTORS THOUGHT THE MEN HAD BEEN POISONED, BUT COULDN'T WORK OUT WHAT WAS WRONG WITH THEM.	18	23	1	18	23	0	0.284
8	THE REASON THE THREE MEN WERE TAKEN TO THE HOSPITAL IS BECAUSE THE PUFFER FISH IS ALSO VERY POISONOUS.	19	29	1	19	29	0	0.295
9	YOUR BRAIN STILL WORKS PERFECTLY HOWEVER, SO YOU KNOW YOU ARE DYING, BUT YOU CAN'T SPEAK OR DO ANYTHING ABOUT IT.	21	28	3	7	9.3333	2	0.289
10	MOST PEOPLE WHO DIE FROM EATING FUGU THESE DAYS ARE PEOPLE WHO HAVE TRIED THEIR HAND AT PREPARING THE FISH THEMSELVES.	21	28	1	21	28	1	0.257
	Mean score	16.2	24.4	1.3	13.9	21.2	1.2	0.278

Three sentences were selected out of the original 55 sentences by greedy search (full search) so that correlation between the DNN-GOP scores and manual ratings of the selected 10 sentences could be maximized. The top five combinations of three sentences are listed up in Table 2. The correlation values of the top five combinations ranged from 0.803 to 0.793, all of which outperformed the correlation (r=0.738, p<.01) between human ratings and DNN- based GOP scores of the 10 sentences from the teacher.

Based on the comparison of Table 1 with Table 2, only one sentence "THE BOY SAID THAT IT HAD ALREADY BEEN BROKEN BEFORE HE AND HIS FRIEND WENT TO THE HOUSE." was identical. Just one out of the 10 sentences selected by the teacher was included in the top five combinations of three sentences for maximizing correlation. The same tendencies were observed in the top 10 and 100 combinations of three sentences. As just described, it was found that the number of shared sentences between teacher-selected sentences and machine-suggested ones was much smaller than expected, and thus the teacher's strategies in selecting stimulus sentences turned out to be inappropriate in terms of maximizing correlation.

7. Characteristics of machine-suggested sentences

7.1 Tendencies in machine-suggested sentence sets

Examining sentence sets suggested by the machine showed that some particular sentences frequently appeared in the sets and seemed to have something to do with maximizing correlation. In Table 2, for instance, the sentence beginning with "HELLO CAROL" showed up five times, and the sentence starting with "THE POLICE OFFICER ASKED" appeared three times. The same tendency was observed in the bottom (worst) five combinations of three sentences for maximizing correlation as shown in Table 3. The sentence "HI, MY NAME IS AKIRA." showed up five times. The reason why "HELLO CAROL. I SAW YOUR HOMEPAGE AND LIKE IT A LOT. YOUR PHOTO WAS REALLY SOMETHING." was treated as one unit in Sentence 1 in Table 2 is that these three were presented to participants at one time in a shadowing session, and they were required to shadow these three utterances all together. On the other hand, "HI, MY NAME IS AKIRA." in Sentence 1 in Table 3 was presented to the participants, and they were asked to shadow only this sentence at a given time.

Table 2: *Top 5 combinations of three sentences to maximize correlation*

	Top five combinations of three sentences for maximazing correlation			sentence length		sentence complexity	T-unit length		prosody	automatic score	Correlation coefficient
	Sentence 1	Sentence 2	Sentence 3	word number	syllable number	T-unit	word number	syllable number	liaison	DNN-GOP	
1	HELLO CAROL. I SAW YOUR HOMEPAGE AND LIKE IT A LOT. YOUR PHOTO WAS REALLY SOMETHING.	THE POLICE OFFICER ASKED, "WHY WERE YOUR FINGERPRINTS FOUND ALL OVER THE DOOR?"	DESPITE THE DANGER OF FUGU POISONING, THIS STRANGE, UGLY AND VERY POISONOUS FISH IS ACTUALLY A VERY EXPENSIVE AND VERY POPULAR KIND OF FOOD IN JAPAN.	18.3	28.3	1.7	14.8	23.4	0.3	0.288	0.803
2	HELLO CAROL. I SAW YOUR HOMEPAGE AND LIKE IT A LOT. YOUR PHOTO WAS REALLY SOMETHING.	THE BOY SAID THAT IT HAD ALREADY BEEN BROKEN BEFORE HE AND HIS FRIEND WENT TO THE HOUSE.	GLOVES, CHOCOLATES AND EVEN UNDERWEAR HAVE ALL BEEN POPULAR AS GIFTS.	15.0	20.0	1.7	11.4	15.1	1.3	0.303	0.797
3	HELLO CAROL. I SAW YOUR HOMEPAGE AND LIKE IT A LOT. YOUR PHOTO WAS REALLY SOMETHING.	GLOVES, CHOCOLATES AND EVEN UNDERWEAR HAVE ALL BEEN POPULAR AS GIFTS.	THE THREE MEN HAD A CLOSE CALL BUT THEY ALL SURVIVED.	12.7	17.3	2.0	7.3	10.4	1.3	0.309	0.795
4	HELLO CAROL. I SAW YOUR HOMEPAGE AND LIKE IT A LOT. YOUR PHOTO WAS REALLY SOMETHING.	THE POLICE OFFICER ASKED, "WHY WERE YOUR FINGERPRINTS FOUND ALL OVER THE DOOR?"	THE THREE MEN HAD A CLOSE CALL BUT THEY ALL SURVIVED.	13.3	17.7	2.0	7.9	10.8	1.0	0.303	0.794
5	HELLO CAROL. I SAW YOUR HOMEPAGE AND LIKE IT A LOT. YOUR PHOTO WAS REALLY SOMETHING.	THE BOY SAID THAT IT HAD ALREADY BEEN BROKEN BEFORE HE AND HIS FRIEND WENT TO THE HOUSE.	THE POLICE OFFICER ASKED, "WHY WERE YOUR FINGERPRINTS FOUND ALL OVER THE DOOR?"	15.7	20.3	1.7	12.1	15.4	1.0	0.297	0.793

Table 3: *Bottom (Worst) 5 combinations of three sentences to maximize correlation*

	Bottom (Worst) five combinations of three sentences for maximazing correlation			sentence length		sentence complexity	T-unit length		prosody	automatic score	Correlation coefficient
	Sentence 1	Sentence 2	Sentence 3	word number	syllable number	T-unit	word number	syllable number	liaison	DNN-GOP	
1	HI, MY NAME IS AKIRA.	HE FINDS A BOY STANDING NEARBY.	WHENEVER IT IS ATTACKED, THE FISH PUFFS UP ITS BODY TO OVER TWICE ITS NORMAL SIZE.	9.0	11.3	1.0	9.0	11.33	0.3	0.358	0.426
2	HI, MY NAME IS AKIRA.	WHENEVER IT IS ATTACKED, THE FISH PUFFS UP ITS BODY TO OVER TWICE ITS NORMAL SIZE.	FUGU IS SAID TO BE SO DELICIOUS THAT IT HAS EVEN STARTED TO BE IMPORTED INTO HONG KONG AND THE UNITED STATES.	14.3	19.3	1.0	14.3	19.33	0.3	0.360	0.432
3	HI, MY NAME IS AKIRA.	HE FINDS A BOY STANDING NEARBY.	THEY COULDN'T SPEAK, AND THEY HAD TROUBLE BREATHING.	6.3	8.7	1.0	6.3	8.67	0.3	0.339	0.441
4	HI, MY NAME IS AKIRA.	WHENEVER IT IS ATTACKED, THE FISH PUFFS UP ITS BODY TO OVER TWICE ITS NORMAL SIZE	MOST PEOPLE WHO DIE FROM EATING FUGU THESE DAYS ARE PEOPLE WHO HAVE TRIED THEIR HAND AT PREPARING THE FISH THEMSELVES.	14.0	18.0	1.0	14.0	18.00	0.3	0.343	0.446
5	HI, MY NAME IS AKIRA.	HE FINDS A BOY STANDING NEARBY.	HE WANTS TO KNOW HOW THE DOOR OF THE MACDONALD'S HOUSE WAS BROKEN OPEN.	8.3	10.3	1.0	8.3	10.33	0.3	0.342	0.447

7.2 Three attributes of selected sentences

To investigate what kind of characteristics are objectively observed in sentences forming combinations for maximizing correlation, three attributes were added for each of the original 55 sentences: sentence length, sentence complexity and number of liaisons.

Sentence length was calculated by the number of words and syllables in the sentence. Sentence complexity was measured by T-unit length. T-unit (minimal terminable unit) is defined as a main clause plus all subordinate clauses and nonclausal structures attached to or embedded in it [16-17]. T-unit length calculated by the number of words or syllables per T-unit has been reported to indicate sentence complexity and has been widely used in research of applied linguistics to date [18-19]. One T-unit contains only one main clause, which is counted as one. No matter how many subordinate clauses may follow a main clause, none of the subordinate clauses are counted. For example, the number of T-units in sentence A is one, and its T-unit length is four. The number of T-units in sentence B is one and its T-unit length is ten. The number of T-units in sentence C is two, because two main clauses are connected by the co-conjunction "and." As a result, the number of T-units in sentence C is two, and its T-unit length is five (ten divided by two).

- Sentence A: Her eyes are blue.
- Sentence B: When the movie is over, we are going straight home.
- Sentence C: The police officer blew his whistle and the truck stopped.

7.3 Comparison of top and bottom combinations of three sentences

The number of possible combinations in choosing three sentences out of the 55 original sentences is 26,235. To observe how the three attributes mentioned above are different between the top- and bottom-ranked combinations of three sentences, the top10, 100 and 1000 and the bottom 10, 100 and 1000 combinations were listed up, and the mean scores of their three attributes are shown in Table 4.

8. Discussion

In this study two correlations were compared: correlation A between human ratings and DNN-based GOP scores of the 10 sentences selected by the teacher and correlation B between human ratings of the 10-sentence set by the teacher and DNN-based GOP scores of three-sentence combinations chosen out of the original 55 sentences by greedy search (full search) for maximizing correlation. Since correlation A was calculated based on the same 10-sentence set, this can be regarded as a closed result. On the other hand, correlation B was computed based on two different groups: the teacher-selected 10-sentence set and the three-sentence sets chosen by greedy search (full search). The number of combinations of three sentences is 26,235. In that sense, correlation B can be thought of as a partially open result compared to correlation A.

The correlation values in the open result in Table 2 outperformed the closed result (correlation value, r=0.738, p<.01). This implies that machine-suggested sentence sets could be more suitable for automatic assessment of L2 shadowing than teacher-selected sentence sets.

To investigate what kind of sentence combinations maximizes/minimizes the correlation, 10, 100 and 1000 combinations were chosen from the top and bottom. To analyze some characteristics of selected sentences, three criteria were set up: sentence length, sentence complexity and number of liaisons. The mean scores of each attribute were calculated across seven groups: top 10, 100 and 1000 combinations, bottom (worst) 1000, 100 and 10 combinations and 10 teacher-selected sentences as in Table 4.

Table 4: *Mean scores of attributes of selected sentences between top and bottom (worst) 10, 100 and 1000 combinations*

		sentence length		sentence complexity	T-unit length		prosody	automatic score
		word number	syllable number	T-unit	word number	syllable number	liaison	DNN-GOP
Top	10 Mean	14.8	20.8	1.7	10.8	15.5	0.9	0.298
	STD	1.5	2.9	0.1	2.0	3.4	0.4	0.009
	100 Mean	14.5	21.2	1.6	11.1	16.6	0.8	0.295
	STD	2.0	3.3	0.2	2.6	3.9	0.4	0.010
	1000 Mean	14.0	20.8	1.4	11.8	17.8	0.8	0.293
	STD	2.3	3.7	0.4	2.8	4.5	0.4	0.012
Bottom (Worst)	1000 Mean	10.5	14.8	1.1	10.0	14.1	0.5	0.329
	STD	2.6	4.2	0.2	2.7	4.4	0.4	0.019
	100 Mean	10.2	13.9	1.0	9.9	13.6	0.4	0.341
	STD	2.8	4.2	0.1	3.0	4.4	0.3	0.014
	10 Mean	9.8	12.9	1.1	9.5	12.4	0.4	0.343
	STD	3.5	4.9	0.2	3.8	5.2	0.2	0.009
10 teacher-selected sentences	Mean	16.2	24.4	1.3	13.9	21.2	1.2	0.287

In the top three groups, average scores of sentence length calculated by the number of words were 14.8, 14.5 and 14.0, respectively. Mean scores of sentence length by the number of syllables were 20.8, 21.2 and 20.8, respectively. On the other hand, in the bottom three groups, average scores of sentence length by the number of words were 10.5, 10.2 and 9.8, respectively. These results show that three-sentence combinations consisting of shorter sentences led to lower correlation in the open result. In other words, if stimulus sentences are too short, they are too easy for participants to shadow, and thus they are not suitable for automatic assessment.

To check how well the participants shadowed across the seven groups, the DNN-based GOP scores were calculated. Table 4 shows that the DNN-GOP scores in the three bottom (worst) groups were 0.329, 0.341 and 0.343, respectively, whereas, those in the three top groups were 0.298, 0.295 and 0.293. The higher the DNN-GOP scores were, the better the participants shadowed. Therefore, it could be said that sentences consisting of around 14 words were more suitable as stimulus sentences in these shadowing tasks than those of around 10 words, based on the comparison of sentence length between the top and bottom three combinations in Tables 2, 3 and 4.

The average sentence length of the 10 teacher-selected sentences was 16.2, which is longer than the mean sentence length in the top three groups (around 14) in Table 4. This implies that the teacher-selected sentences were so difficult that participants did not do well in those sentences, and consequently the number of shared sentences between teacher-selected sentences and machine-suggested ones was much smaller than expected.

In the same procedure above, other attributes were calculated and analyzed as follows (Table 4):
The average score of sentence length by the number of syllables in the top three groups was around 20. That in the bottom three groups was around 14. That in the 10 teacher-selected sentences was 24.4.

The mean T-unit score indicating the number of main clauses in the top three groups was around 1.6. That in the bottom (worst) three groups was around 1.1. That in the teacher-selected 10 sentences was 1.3.

T-unit length was calculated by the total number of words or syllables divided by the number of T-units included in a sentence. The average scores of T-unit length by the number of words and those by the number of syllables in the top three groups were around 11 and 16, respectively. Those in the bottom three groups were around 10 and 13, respectively. Those in the 10 teacher-selected sentences were 13.9 and 21.2, respectively.

The average number of liaisons included in a sentence in the top three groups was around 0.8. That in the bottom (worst) three groups was around 0.4. That in the teacher-selected 10 sentences was 1.2.

The average DNN-based GOP scores in the top three groups were around 0.29. Those in the bottom three groups were around 0.34. That in the teacher-selected 10 sentences was 0.287.

Considering these results, sentences included in the bottom groups were easy or too easy for participants to shadow. On the other hand, the 10 teacher-selected sentences were difficult or too difficult for L2 shadowing.

9. Conclusions and future studies

Based on this investigation it could be concluded that stimulus sentence combinations for automatic assessment of L2 shadowing should not be too easy or too difficult. Stimulus sentences should possess moderate complexity in terms of the number of words and syllables and sentence construction complexity measured by T-unit length. These are thought to be required conditions to improve reliability in automatic assessment of L2 shadowing.

Although this conclusion is not so surprising, the results of this investigation might deepen and broaden the insights of teachers and educators into selecting proper materials for assessment. In many cases teachers tend to choose syntactically and semantically difficult sentences as stimuli. Teachers might have the preconception that in L2 shadowing more difficult sentences tend to have more power to discriminate learners' proficiency, lead to learners' best performances and reveal what they can do using their utmost linguistic abilities. However, the results of this investigation did not support such a preconceived idea.

In language education stimulus sentences are usually chosen and determined based on veteran teachers' knowledge and experiences. It is very rare that stimulus sentences are selected systematically by a machine and compared to teacher-selected sentences. The results of this investigation were obtained by objective and systematic analysis, which has rarely been conducted in language education and might shed light on altering teachers' beliefs in assessment.

L2 shadowing is thought to be a very cognitively demanding task, because it requires learners to do two things simultaneously: listening comprehension and oral reproduction. Therefore, L2 shadowing itself might be difficult enough to differentiate learners' proficiency, and thus stimuli should not be so difficult.

As for future studies it is very crucial to conduct the same experiment using passages different from those used in this study and examine whether the results obtained from this investigation are dependent on the given sentence sets or not. In other words, if stimulus sentences are selected from different passages to meet the required conditions proposed by the present study, will the assessment result in the same way as in this investigation? After this probing work, conditions obtained from this study can be considered as sufficient conditions as well as necessary ones, and the conclusions will be more generalized.

10. Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers JP16H03084, JP16H03447, and JP26240022. We'd like to thank all students and educators who participated in this experiment, including Prof. Kay Husky (Tokyo International University) for her professional advice on this paper.

11. References

[1] W.D.Marslen-Wilson, "Speech shadowing and speech comprehension. *Speech Communication*," vol. 4, nos.1-3, pp.55-73, 1985.

[2] H. Mitterer and M. Ernestus, "The link between speech perception and production is phonological and abstract: Evidence from the shadowing task," *Cognition*, vol.109, no.1, pp.168-173, 2008.

[3] P. W. Carey, "Verbal retention after shadowing and after listening," *Perception & Psychophysics*, vol.9, no.1, pp.79-83, 1971.

[4] S. Miyake, "Cognitive processes in phrase shadowing and EFL Listening," *JACET (Japan Association of College English Teachers) Bulletin*, vol.48, pp.15-28, 2009.
[5] Y. Hamada, "The effectiveness of pre- and post-shadowing in improving listening comprehension skills," *The Language Teacher*, vol.38, no.1, pp.3-10, 2014.
[6] Y. Hamada, "Shadowing: Who benefits and how? Uncovering a booming EFL teaching technique for listening comprehension," *Language Teaching Research*, vol.20, no.1, pp.35-52, 2016.
[7] T. Hori, "Exploring Shadowing as a Method of English Pronunciation Training," A Doctoral Dissertation Presented to the Graduate School of Language Communication and Culture, Kwansei Gakuin University, 2008.
[8] K. T. Hsieh, D. H. Dong, and L. Y. Wang, "A preliminary study of applying shadowing technique to English intonation instruction," *Taiwan Journal of Linguistics*, vol. 11, no. 2, pp. 43-65, 2013.
[9] D. Luo, N. Minematsu, Y. Yamauchi, and K. Hirose, "Automatic assessment of language proficiency through shadowing," in ISCSLP'08, *– 6th International Symposium on Chinese Spoken Language Processings*, pp.1-4, 2008.
[10] K. Tamai, *A study on the effectiveness of shadowing as an instructional method of listening*. Tokyo: Kazama Shobou, 2005.
[11] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, no. 2, pp. 95-108, 2000.
[12] D. Luo, N.Shimomura, N.Minematsu, Y.Yamauchi, and K.Hirose, "Automatic pronunciation evaluation of language learners' utterances generated through shadowing," *Proceedings of INTERSPEECH 2008*, pp.2807-2810, 2008.
[13] D. Luo, N. Minematsu, Y. Yamauchi, and K. Hirose, "Analysis and comparison of automatic language proficiency assessment between shadowed sentences and read sentences." in *SLaTE* 2009, pp. 37-40, 2009.
[14] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A.Senior, V.Vanhoucke, P.Nguyen, T.N.Sainath, & B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82-97, 2012.
[15] J. Yue, F. Shiozawa, S. Toyama, Y. Yamauchi, K. Ito, D. Saito, N. Minematsu, "Automatic scoring of shadowing speech based on DNN posteriors and their DTW," in *INTERSPEECH*, 2017. (accepted)
[16] K. W. Hunt, *Grammatical structures written at three grade levels*, Research Report no.3, Urbana, IL: National Council of Teachers of English, 1965.
[17] K. W. Hunt, "Syntactic maturity in school children and adults," *Monographs of the Society for Research in Child Development*, vol. 35, no.1, pp.1-67, 1970.
[18] K. Hirano, "Research on T-unit measures in ESL," *Bulletin of Joetsu University of Education*, vol.8, sect.2, pp.67-77, 1989.
[19] C. G. Polio, "Measures of linguistic accuracy in second language writing research," *Language Learning*, vol.47, no.1, pp.101-143, 1997.