Vol.2018-SLP-125 No.16
2018/12/10

Influence of content variations on native speakers' fluency of shadowing

Tasavat Trisitichoke[1,a)] Shintaro Ando[1,b)] Yusuke Inoue[1,c)] Daisuke Saito[1,d)]
Nobuaki Minematsu[1,e)]

Abstract: In our recent studies concerning speech shadowing, a new term of *shadowability* was introduced to indicate how smoothly listeners can shadow given utterances. In this study, to interpret shadowability experimentally, various kinds of spoken Japanese utterances, read aloud by an expert narrator, are used as stimuli and presented to native listeners who are asked to shadow them. The stimuli sentences are designed by controlling comprehensibility qualitatively, in order to examine how shadowability changes due to semantic/syntactic contents of a given utterance. It is shown that qualitatively controlled comprehensibility of stimuli strongly influences shadowers' performances.

Keywords: speech shadowing, natives' shadowing, intelligibility, comprehensibility, shadowability, second language learning, syntactic/semantic content

1. Introduction

In our recent studies [1, 2], L2 learners' utterances were shadowed by native listeners in order to objectively measure comprehensibility of the L2 utterances. Analysis of the shadowing utterances showed that, when considering comprehensibility of learners' utterances, natives' shadowings are more informative indicators than learners' utterances. Although analysis of learners' utterances can tell how similar they are to native pronunciation, it is much more effective to turn to natives' responsive shadowings, when learners are more interested in comprehensibility of their pronunciation.

[1, 2] also coined a new term of *shadowability*, indicating how smoothly listeners can shadow given utterances. Due to high cognitive load imposed on listeners in shadowing, it was discussed theoretically that shadowability can be interpreted to be closer to comprehensibility than to intelligibility of utterances. However, since native listeners' oral repetition tests can define intelligibility of utterances objectively [3, 4], shadowability can also be simply interpreted as *online* intelligibility.

To interpret shadowability experimentally, in this paper, various kinds of spoken Japanese, given from a professional narrater, are used as stimuli and presented to native listeners who are asked to shadow them. The stimuli are designed by controlling comprehensibility of content qualitatively.

Results show that two kinds of shadowability scores, accuracy of articulation and delay of shadowing, which are calculated automatically using speech technologies, are strongly influenced by comprehensibility of the stimuli. It can be concluded that shadowability is correlated with comprehensibility, although its measurement method can superficially characterize it as online intelligibility.

[1] The University of Tokyo, Bunkyō, Tokyo 113-8654, Japan
[a)] tasavat@gavo.t.u-tokyo.ac.jp
[b)] s_ando@gavo.t.u-tokyo.ac.jp
[c)] inoue0124@gavo.t.u-tokyo.ac.jp
[d)] dsk_saito@gavo.t.u-tokyo.ac.jp
[e)] mine@gavo.t.u-tokyo.ac.jp

2. Three types of measurement

2.1 Intelligibility and comprehensibility

In applied linguistics, intelligibility and comprehensibility are defined differently [5]. Intelligibility indicates, for a given utterance, how accurately linguistic units such as words can be identified. Degree of intelligibility of a given utterance can be measured objectively, for example, by asking native listeners to repeat or write down that utterance. Correct identification rate can represent intelligibility of that utterance.

Comprehensibility of an utterance, on the other hand, indicates on how easily and smoothly listeners can understand the content of the utterance. We can point out cognitive load or listening efforts as semantically similar terms. Comprehensibility is often quantified using subjective questionnaires or comprehension tests imposed on listeners. Since correct comprehension often requires syntactic analysis and pragmatic analysis in addition to correct identification of words, it can be considered that comprehensibility covers and extends beyond intelligibility. In our study, we are aiming at automatic and objective measurement of comprehensibility of given L2 utterances.

2.2 Shadowability

[1, 2] quantified shadowability by two aspects of shadowing utterances, one is related to accuracy of articulation and the other is to delay of shadowing.

For accuracy of articulation, GOP (Goodness Of Pronuncia-

Table 1 Various contents used for shadowing

set	source
A	a very famous classical tale (Momotarō)
B	easy articles from NHK NWE*
C	random word sequences from NHK NWE*
D	original articles from NHK News Web
E	articles from Nikkei Science
F	random concatenation of Japanese characters

*NWE(News Web Easy): a Japanese news site for foreigners who are learning Japanese (`https://www3.nhk.or.jp/news/easy/`).

Table 2 Comparison of the six stimulus sets

set	WF	CWP	CPP	CSS
A	○	◎	◎	◎
B	◎	◎	○	◎
C	◎	×	×	×
D	○	○	○	○
E	△	○	○	○
F	×	×	×	×

WF: word frequency
CWP: cross-word predictability
CPP: cross-phrase predictability
CSS: complexity of syntactic structure

Table 3 The number of GOPs for stimulus sets

set	**A**	**B**	**C**	**D**	**E**	**F**
number	15	16	20	18	7	15

tion) is adopted as it is widely used as a baseline feature to indicate accuracy of articulation. GOP is theoretically defined as posterior $P(c_i|o_t)$, where o_t is a speech feature at time t, and c_i is phonemic class i intended by a speaker. In [2], after forced alignment, GOP was calculated for each phonemic unit, and utterance-unit GOP was calculated by averaging the phoneme-unit GOP scores of an utterance.

As for delay of shadowing, by comparing forced alignment of a presented utterance and that of its corresponding shadowing, the temporal gap between every pair of phoneme boundaries is obtained between the two utterances. The phoneme-based temporal gaps obtained from the two were averaged to define delay of shadowing between the two utterances. Shadowing is often performed with delay of approximately 1 second to a presented utterance.

For detailed procedures of training DNN-based acoustic models and calculating the two kinds of scores, readers should refer to [1] and [2].

3. Experiments

3.1 Various contents for shadowing

To analyze the influence of linguistic content on natives' shadowing, six sets of readings, as shown in Table 1, were prepared as stimuli.

Easy-to-understand sentences were collected from a famous classical tale, Momotarō (**A**) and from NHK News Web Easy (**B**), which is a content provided for foreigners learning Japanese. Highly intelligible but extremely incomprehensible stimuli were prepared by randomly concatenating content words found in NWE (**C**). Seemingly rather difficult-to-understand sentences were collected from science magazines Nikkei Science (**E**).

As reference, random concatenations of Japanese characters (Hiragana) were also used as stimuli (**F**). Prosodic control for reading these random sequences of Hiraganas was done by simulating that in Momotarō (**A**). In other words, set **F** was prepared by replacing each Hiragana in Momotarō with another. Here, so-called Seion (清音) was used exclusively for replacement.

Very subjective and qualitative comparison of these six sets of stimuli is done in Table 2. Four linguistic factors are considered to control comprehensibility of the reading stimuli. They are word frequency (word familiarity[*1]), cross-word predictability, cross-phrase or cross-sentence predictability, and complexity of syntactic structure.

Each set is composed of twenty utterances, each of which is composed of a sentence or some phrases. These utterances were given by a professional female narrator to ensure smoothness of speech production even in the case of set **F**.

3.2 Subjects

Seven adult subjects, five males and two females, participated in the experiments. The male subjects are university students majoring in engineering and word familiarity of set **E** will be high to them. The female subjects are secretaries who did not major in engineering or science and thus word familiarity of some technical terms in set **E** will be lower.

3.3 Procedures

Each set of twenty utterances are divided into four groups of five utterances in each. In total, we have 24 groups. Using these groups, the shadowing experiments were carried out in a particular manner. Firstly, to provide an overall picture for subjects, one group from each set (**A** to **F**) was presented consecutively. Then, the remaining 18 groups were randomly selected and presented.

After a simple shadowing practice, subjects were asked to shadow all the 120 utterances, where they were not allowed to repeat shadowing any given utterance unless considered necessary.

3.4 Analysis

When shadowing a given utterance, if several pauses are found in the utterance, shadowing becomes easy, arguably due to usage of short-term memory. For fair comparison among the six stimulus sets, phrases that are longer than or equal to 10 morae, and are read aloud with no pause, were manually selected for analysis.

Furthermore, not a small number of phrases in set **E** are composed of ordinary words only, not including any scientific or technical terms. So, oral phrases including those terms that require high-school science knowledge were selected manually.

Analysis was done only for these selected phrases. Table 3 shows the number of utterances available.

A GOP score and a delay is calculated from a shadowing of

[*1] In Momotarō, words, phrases, even sentences are highly predictable, but some phrases are used in daily conversation very rarely, such as 芝刈りに行く.

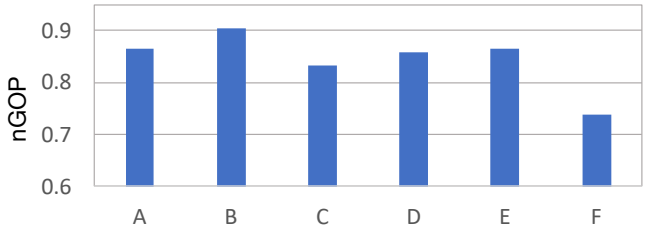

Fig. 1 nGOP scores for the six stimulus sets

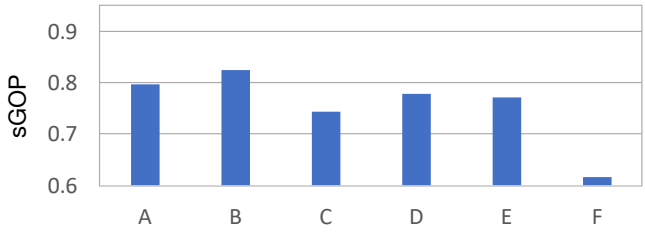

Fig. 2 sGOP scores for the six stimulus sets

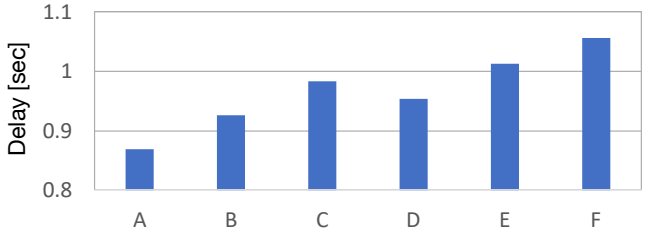

Fig. 3 Delay of shadowing for the six stimulus sets

each utterance. GOP scores were calculated separately for each set from the professional narrator (nGOP) and from the subjects (sGOP). In the latter case, for each utterance, the highest and the lowest GOP scores were removed.

T-tests were done for both GOP scores and delay to examine between sets whether significant differences at 5% are found.

3.5 Results and discussion

Figures 1 and 2 show the GOP scores for utterances of the narrator and those for shadowings. Significant differences observed are also listed in Table 4 and Table 5. nGOP and sGOP show a very similar variation and its pattern is considered to be very reasonable due to the linguistic contents of the stimuli (see Table 2). We consider it interesting that GOP of even a professional reading is influenced by its content.

Since **B** and **D** are from news articles, we consider that they are more ordinary and typical compared to the other sets. Table 4 and Table 5 show that, for nGOP, **B** and **D** have significant differences to **ACDF** and **BCF**, respectively, and that, for sGOP, **B** and **D** have significant differences to **CDEF** and **BCF**, respectively.

Words used in set **C** are very easy but they do not have any meaning as sentence. We can say that set **C** are totally intelligible but totally incomprehensible. For both nGOP and sGOP, their GOP scores are significantly different from those in **B** and **D**. Further, between **B** and **D**, they are also different both for nGOP and sGOP. From these results, it can be concluded that comprehensibility of the stimuli strongly influences GOP scores.

Delays of shadowing are shown in Figure 3. Delays in set **A** (Momotarō) are the smallest because every single word is highly predictable. T-tests' results for delays of shadowing of **B** and **D** (see Table 6) showed that significant differences are found to **ACEF** and **AEF**, respectively. **C** is judged to be significantly different only from **B**. In this case, significant differences are not found between **B** and **D**. Comparing the results of GOP and those of delay, significant differences are found in different stimulus pairs. However, we consider that it is adequate to claim that comprehensibility of the stimuli also influences delay of shadowing.

Table 4 Significant differences for nGOP scores

set	A	B	C	D	E	F
A		○	○			○
B	○		○	○		○
C	○	○		○		○
D		○	○			○
E						○
F	○	○	○	○	○	

Table 5 Significant differences for sGOP scores

set	A	B	C	D	E	F
A			○			○
B			○	○	○	○
C	○	○		○		○
D		○	○			○
E		○				○
F	○	○	○	○	○	

Table 6 Significant differences for delay

set	A	B	C	D	E	F
A		○	○	○	○	○
B	○		○		○	○
C	○	○				○
D	○				○	○
E	○	○		○		
F	○	○	○	○		

4. Conclusions

In this study, it was shown experimentally that qualitatively controlled comprehensibility of stimuli strongly influenced shadowers' performances. Interestingly enough, reading performances of a professional narrator were also influenced by comprehensibility of given text, even after careful recording rehearsals. The authors consider that GOP and delay, which are automatically calculated from natives' shadowings, are very helpful to predict comprehensibility of learners' utterances.

Acknowledgments This work was financially supported by JSPS or MEXT KAKENHI JP26118002 and JP18H04107.

References

[1] Y. Inoue *et al.*, "A study of objective measurement of comprehensibility through native speakers' shadowing of learners' utterances," *Proc. INTERSPEECH*, 1651–1655, 2018.
[2] 井上他，秋季音講論，2-P-11，2018.
[3] J. Bernstein, *Proc. ICPhS*, 1581–1584, 2003.
[4] N. Minematsu, *et al.*, *Proc. INTERSPEECH*, 1481–1484, 2011.
[5] M. J. Munro and T. M. Derwing, *Language Learning*, 45, 1, 73–97, 1995.