Influence of content variations on native speakers' performance of shadowing*

☆ Tasavat Trisitichoke, Shintaro Ando, Yusuke Inoue, Daisuke Saito, Nobuaki Minematsu (The University of Tokyo)

1 Introduction

In our recent studies [1, 2], to objectively measure comprehensibility of learners' utterances, they were shadowed by native listeners. Analysis of the shadowing utterances showed that natives' shadowings are more informative than learners' utterances to predict comprehensibility of learners' utterances. Although analysis of learners' utterances can tell how similar they are to native pronunciation, when learners want to know how comprehensible their utterances are, it will be much more valid to turn to natives' responsive shadowings.

In [1, 2], a new term of *shadowability* was introduced to indicate how smoothly listeners can shadow given utterances. Due to high cognitive load imposed on listeners in shadowing, we 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 be simply interpreted as *online* intelligibility. In this paper, to interpret shadowability experimentally, various kinds of spoken Japanese are used as stimuli and presented to native listeners who are asked to shadow them. The stimuli are designed by controlling comprehensibility and given from a professional narrater.

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

2 Three measures of abilities

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 that utterance. Correct identification rate can represent intelligibility of that utterance. Comprehensibility of an utterance means how easily and smoothly listeners can understand the content of that utterance, 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, the authors consider that comprehensibility covers intelligibility and represents more.

2.2 Shadowability

In [1, 2], we quantified shadowability by focusing on 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 Pronunciation) is adopted because it is widely used as baseline feature to indicate accuracy of articulation. GOP is theoretically defined as posterior $P(c_t^i|o_t)$, where o_t is a speech feature at time t, and c_t^i is phonemic class i intended at time t 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].

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 means News Web Easy and it is a Japanese news site for foreigners who are learning Japanese.

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, SS: complexity of syntactic structure

3 Experiments

3.1 Various contents for shadowing

To analyze the influence of linguistic content on natives' shadowing, six sets of readings were prepared as stimuli, shown in Table 1. Easy-to-understand sentences were collected from a famous classical tale, Momotarō (**A**) and NHK News Web Easy (**B**), which is 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 (**E**). As reference, random concatenations of Japanese characters (Hi-

*提示内容の多様性が母語話者シャドーイングに与える影響，トリシティショーク・タサバット, 安藤慎太郎, 井上雄介, 齋藤大輔, 峯松信明（東京大学）

Table 3: The number of GOPs for stimulus sets

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

ragana) were also used as stimuli (**F**). Prosodic control for reading these random sequences of Hiraganas was done by simulating that in Momotarō. 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 word familiarity of some technical terms in set **E** will be lower.

3.3 Procedures

Each set has twenty utterances and they are divided into four groups, each of which has five utterances. 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 set **A** to set **F** was presented consecutively. Then, the remaining 18 groups were randomly selected and presented.

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

3.4 Results and discussion

When shadowing a given utterance, if several pauses are found in that utterance, shadowing becomes easy. For fair comparison among the six stimulus sets, we manually collected phrases for analysis that are longer than or equal to 10 morae and were produced orally by the professional narrator with no pause. Further, not a small number of phrases in set **E** are composed of ordinary words, 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 on shadowings for these selected phrases. Table 3 shows the number of utterances available. A GOP score and a delay is calculated from a shadowing of each utterance.

GOP scores were calculated separately for each set from the professional narrator (nGOP) and from the seven subjects (sGOP). In the latter case, for each utterance, the highest and the lowest GOP scores were removed for stable analysis. T-tests were done for these GOP scores to examine between which sets,

[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.

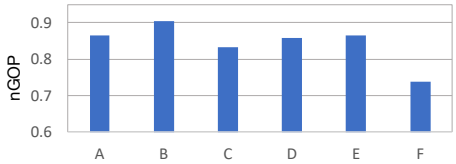

Fig. 1: nGOP scores for the six stimulus sets

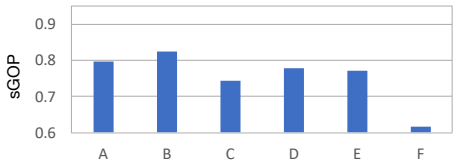

Fig. 2: sGOP scores for the six stimulus sets

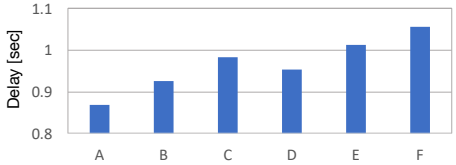

Fig. 3: Delay of shadowing for the six stimulus sets

significant differences at 5% are found. Figures 1 and 2 show the results and a very similar GOP variation is found between them. The variation pattern is considered to be very reasonable due to the linguistic contents of the stimuli (see Table 2). It is interesting that GOP of even a professional reading is influenced by its content. T-tests' results of **B** and **D** showed that, for nGOP, significant differences are found to **ACDF** and **BCF**, respectively and, for sGOP, they are found to **CDEF** and **BCF**, respectively. Set **C** are intelligible but incomprehensible stimuli and their GOP scores are different from those in **B** and **D**. Between **B** and **D**, they are also different. We can say 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 of **B** and **D** 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.

4 Conclusions

It was shown 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 helpful to predict comprehensibility of learners' utterances.

参考文献

[1] Y. Inoue *et al.*, "A study of objective measurement of comprehensibility through native speakers' shadowing of learners' utterances," *Proc. INTERSPEECH*, 2018 (accepted).
[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.