INFLUENCE OF CONTENT VARIATIONS ON SMOOTHNESS OF NATIVE SPEAKERS' REVERSE SHADOWING

Tasavat Trisitichoke, Shintaro Ando, Yusuke Inoue, Daisuke Saito, Nobuaki Minematsu
School of Engineering, The University of Tokyo
{tasavat,s_ando,inoue0124,dsk_saito,mine}@gavo.t.u-tokyo.ac.jp

ABSTRACT

Recently, a novel method of objective measurement of comprehensibility of L2 utterances was proposed. Native listeners were asked to shadow learners' utterances, and from natives' shadowings, 1) accuracy of shadowers' articulation and 2) delay of shadowing were measured acoustically and automatically. These two measurements were found to be highly correlated with comprehensibility perceived by the shadowers. Comprehensibility, or smoothness of understanding is considered to be characterized well by the two measurements because shadowers have to listen, understand, and repeat simultaneously. This paper aims at validating this method from another viewpoint. Here, listeners are asked to shadow *native* utterances with varying levels of lexical, syntactic, semantic, and pragmatic complexity, which can easily influence perceived comprehensibility. Experiments show that the above two measurements depend strongly on the complexity levels. We can claim again that the two automatic measurements can characterize perceived comprehensibility well.

Keywords: Comprehensibility, natives' shadowing, linguistic complexity, GOP, delay of shadowing

1. INTRODUCTION

In phonetic studies of learners' pronunciation, their deviations from native pronunciation are often discussed [11, 19]. However, some types of foreign accents hardly reduce smoothness of communication [2, 13, 14]. The practical goal of pronunciation training is an intelligible-enough pronunciation, not a native-sounding one [2].

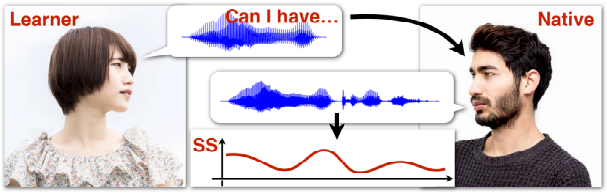
In applied linguistics, intelligibility of an utterance indicates how many linguistic units such as words can be identified correctly. Degree of intelligibility of an utterance can be measured objectively by asking native speakers to transcribe or repeat that utterance after listening to it [1, 13]. Comprehensibility of an utterance means how easily and smoothly listeners can understand the content of that utterance, and degree of comprehensibility has been often quantified by listeners' subjective judgment [13]. Listening effort [18] and cognitive load [4] seem to be strongly related to comprehensibility, i.e. smoothness of understanding.

Intelligibility was measured objectively in [1, 12], where English spoken by immigrants to the USA [1] and English by Japanese college students [12] were presented to native listeners on the telephone. They were asked, after listening, to repeat what they heard. Their repetitions were transcribed manually by technical staff to derive word-based intelligibility of each utterance. However, no good control seems to have been made on listeners' behavior of repetition. Efforts of listening or guessing and delay of repetition depended on listeners and therefore, the obtained intelligibility scores will not always match comprehensibility. Further, the speaking styles observed in the repeated utterances were not discussed.

How to measure objectively comprehensibility, listening effort, or cognitive load when a listener listens to an L2 utterance? Previous studies used physiological sensors for measurement. In [3, 5, 18], EEG (electroencephalogram) recordings were made from listeners and listening efforts were discussed quantitatively and in [4], eye-trackers were used to measure the size of pupils to predict the magnitude of cognitive load when listening. In [7, 8], a novel method of predicting comprehensibility of an utterance was proposed, which did not require any special device. Native listeners were asked to *shadow* learners' utterances, where shadowing indicated not imitating accented pronunciations but repeating in their own native pronunciation what was said. From natives' shadowings, accuracy of shadowers' articulation and delay of shadowing were measured acoustically and automatically. These two measurements were found to be highly correlated to comprehensibility subjectively judged by the shadowers.

This paper aims at validating this method from a different viewpoint. When native speakers repeat or transcribe L2 utterances after listening to those utterances, repetitions or transcriptions indicate intelligibility. Our claim is that, when they shadow, smoothness of their shadowings will indicate comprehensibility. Suppose that native speakers are asked to transcribe utterances from audio books for kids and

Figure 1: Native listeners' *reverse* shadowing



SS means smoothness of shadowing.

those from news broadcast, they will give perfect transcriptions to both although smoothness of understanding may be different. When they shadow those utterances, smoothness of shadowing will be different due to differences in lexical, syntactic, semantic, and pragmatic complexity. In other words, intelligibility can be viewed as speech feature that characterizes results of the process of understanding, and comprehensibility represents how smoothly the process is running. This paper verifies that the two acoustic measurements, related to smoothness of shadowing, strongly depend on the linguistic features of presented utterances, which can directly influence smoothness of understanding.

In previous studies on shadowing including those not related to language learning [9, 15, 17], delay of shadowing was used as main speech feature. Our previous studies [7, 8] and this study are novel in that shadowing behaviors are acoustically characterized by accuracy of articulation as well as delay of shadowing. The first feature is automatically calculated by DNN-based speech recognition modules, which will be explained in the following section.

2. NATIVE LISTENERS' REVERSE SHADOWING OF L2 UTTERANCES

In the conventional form of shadowing, native utterances are presented to learners, who have to listen and repeat as simultaneously as possible. In this study, it is native listeners who have to shadow while listening to and understanding learners' utterances. Figure 1 shows native listeners' *reverse* shadowing, where smoothness of shadowing is acoustically measured as two kinds of acoustic features.

As for shadowers' accuracy of articulation, we use Goodness Of Pronunciation (GOP) measure [6, 20, 21]. GOP is a widely-used baseline speech feature in pronunciation assessment studies and, when GOP is applied to an L2 utterance, it represents how similar that utterance is to the model pronunciation in terms of articulation. GOP is theoretically defined as phoneme-based posterior $P(c_i|o_t)$, where o_t is a speech feature observed at time t, and c_i is phonemic class i. In Figure 1, after forced alignment performed on the native shadowing with the string of phonemes intended by the learner, $P(p_t|o_t)$ is averaged over the entire duration of a given phonemic segment, where p_t is the phoneme shadowed at time t. The GOP of a given segment x is calculated as

$$(1) \quad \mathrm{GOP}(x) = \frac{1}{D_x} \sum_{t \in x} P(p_t|o_t),$$

where D_x is the frame-based total duration of x.

In [7, 8], Japanese spoken by Vietnamese learners and native listeners' shadowings were used for analysis, and $P(p_t|o_t)$ was calculated by a DNN-based front end of a Japanese speech recognizer, trained with CSJ [10]-based KALDI [16]. The GOP scores from natives' shadowings were shown to be very highly correlated with comprehensibilities subjectively judged by the native shadowers.

As for delay of shadowing, in [7, 8], by comparing a forced alignment result of a Vietnamese-Japanese (VJ) utterance and that of its native reverse shadowing (RS), the temporal gap between every pair of phoneme boundaries was obtained between the two utterances. The phoneme-based temporal gaps obtained from the two utterances were averaged to define delay of shadowing between the two. Generally speaking, shadowing is performed with a delay of about 1 second to a presented utterance.

In [7, 8], learners' utterances were obtained by asking learners to read aloud some paragraphs from an textbook the linguistic complexity of which was low. It indicates that the degree of accentedness, not the linguistic content of utterances, influenced strongly the above two acoustic measurements. In this paper, unlike [7, 8], native read-aloud utterances are presented to native shadowers. Here, lexical, syntactic, semantic, and pragmatic complexities of the stimuli vary largely. Since these linguistic features can directly influence smoothness of understanding, if the two acoustic measurements are found to strongly depend on these features, it allows us to claim that the method of natives' reverse shadowing observed as the above two acoustic measurements can be validated from a different viewpoint.

3. EXPERIMENTS

3.1. Various contents for shadowing

To analyze the influence of the linguistic features on smoothness of natives' shadowing, six sets of readings were prepared, as shown in Table 1. Very easy-to-understand sentences were collected from a Japanese famous classical tale, Momotarō (**A**), and from NHK News Web Easy (**B**), which is a news content provided for foreigners learning Japanese.

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 of NWE 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: Qualitative comparison of the stimuli

set	WF	CWP	CPP	CSS
A	M	H	H	H
B	H	H	M	H
C	H	—	—	—
D	M	M	M	M
E	L	M	M	M
F	—	—	—	—

WF: word frequency
CWP: cross-word predictability
CPP: cross-phrase predictability
CSS: easiness of syntactic analysis
H, M, L, —: high, middle, low, extremely low

Table 3: #utterances for the six stimulus sets

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

Highly intelligible but extremely incomprehensible stimuli were prepared by randomly concatenating content words found in NWE (**C**). The original articles of **B** were extracted from NHK News Web (**D**). Rather difficult-to-understand sentences were collected from science magazines of 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 (unvoiced consonant syllable) was used exclusively for replacement. Intuitive 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, cross-word predictability, cross-phrase or cross-sentence predictability, and easiness of syntactic analysis. Their abbreviations are used in Table 2.

Each set had twenty utterances and each utterance was composed of a sentence or some phrases. These utterances were given by a professional female narrator to ensure smoothness of speech production. In recording, she was instructed to read sentences naturally but neutrally with a fixed speaking rate except for **F**. Reading rehearsals were allowed if needed.

3.2. Subjects and procedures

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

Each set of twenty utterances were divided into four groups of five utterances in each. In total, we had 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 to the subjects.

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

3.3. Analysis of smoothness of shadowing

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, only the phrases that are longer than or equal to 10 morae and read aloud by the narrator without pausing were manually selected for analysis. In set **E**, not a small number of phrases were 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 of smoothness of shadowing was done only on these selected utterances. Table 3 shows the number of utterances available.

Two kinds of GOP scores were calculated, one is from a shadowing and the other is from a presented utterance given by the narrator. The former is called subjects' GOP (sGOP) and the latter is called narrator's GOP (nGOP). In sGOP, for a given utterance, the highest and the lowest sGOP scores among the seven subjects were removed. Delay of shadowing was calculated from a pair of a shadowing and its corresponding utterance given by the narrator.

T-tests were done for both GOP scores and delay of shadowing to examine between which sets significant differences at 5% are found. For nGOP, the number of samples is shown in Table 3 and for sGOP, the number is samples is five times larger than the number in Table 3. For analysis of delay, the number of samples is the same as that for sGOP.

Table 4: 5% differences in sGOPs

set	A	B	C	D	E	F
A	–		*			*
B		–	*	*	*	*
C	*	*	–	*		*
D		*	*	–		*
E		*			–	*
F	*	*	*	*	*	–

Table 5: 5% differences in delays

set	A	B	C	D	E	F
A	–	*	*	*	*	*
B	*	–	*		*	*
C	*	*	–			*
D	*			–	*	*
E	*	*		*	–	
F	*	*	*	*		–

Table 6: 5% differences in nGOPs

set	A	B	C	D	E	F
A	–	*	*			*
B	*	–	*	*		*
C	*	*	–	*		*
D		*	*	–		*
E					–	*
F	*	*	*	*	*	–

Figure 2: sGOP scores for the six stimulus sets

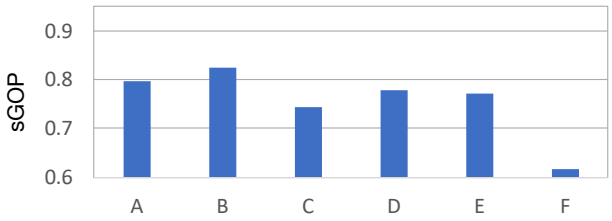


Figure 3: Delay of shadowing for the six sets

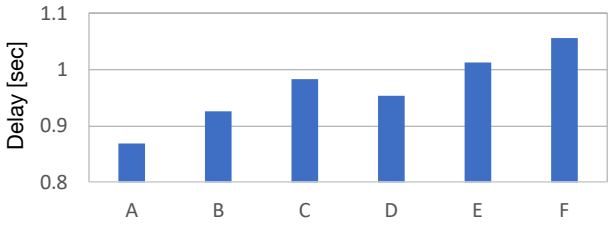
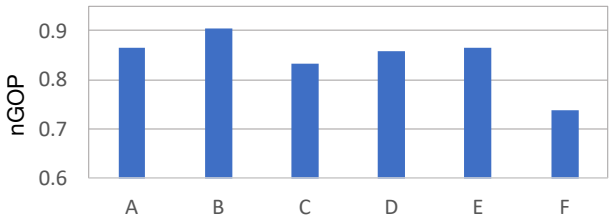


Figure 4: nGOP scores for the six stimulus sets



3.4. Results and discussion

Firstly, we discuss the shadowers' behaviors with sGOP scores and delays. From Table 2, it is expected that **A** and **B** will show high sGOP and short delay, and that **C** and **F** will show low sGOP and long delay. **D** and **E** will show intermediate levels of sGOP and delay. Figures 2 and 3 show averaged sGOP scores and averaged delays of shadowing. From both figures, we can say that the above expectations are almost valid. Furthermore, we can point out that shadowing in **F** is extremely difficult and shadowing in **A** can be done extremely quickly.

Tables 4 and 5 show between which sets significant differences at 5% are found. Since **B** and **D** are from news articles, we consider that they contain more ordinary expressions compared to the other sets. In sGOP, **B** and **D** have significant differences to **CDEF** and **BCF**, respectively. **B** are prepared by editing **D** so that learners of Japanese can understand the meaning. Comprehensibility is different qualitatively between **B** and **D** and sGOP is also different between the two. In delays, **B** and **D** are different from **ACEF** and **AEF**, respectively, meaning no significant delay difference between **B** and **D**.

Words in **C** are very easy to understand but they do not have any meaning as sentence. We can say that words in **C** are totally intelligible but totally incomprehensible. When we see **B** and **C**[1]. **C** is significantly different from **B** both in sGOP and delays.

From these results, we can claim that comprehensibility of utterances, which is controlled in the current experiment not by accentedness of utterances but by their linguistic features, can easily and strongly influence smoothness of shadowing, which is acoustically measured as sGOP and delay.

Next, we focus on nGOP to discuss behaviors of the narrator. Averaged scores of nGOP are shown in Figure 4 and results of t-tests are listed in Table 6. It is very interesting that the nGOP scores show a very similar distribution over the six stimulus sets to the sGOP distribution. Significant differences are found both between **B** and **D** and between **B** and **C**. The professional narrator was asked to read given sentences neutrally at a fixed speaking rate but her reading behaviors were influenced perhaps involuntarily by the linguistic features of given sentences. It is inevitable that shadower's behaviors are influenced by the linguistic features related to comprehensibility.

4. CONCLUSIONS

This study showed that comprehensibility of stimuli strongly influenced not only shadowers' performances but also reading performances of a professional narrator. The experimental results suggest that sGOP and delay, which are calculated acoustically from natives' shadowings, will be very helpful to predict comprehensibility of presented utterances.

This work was supported financially by JSPS KAKENHI JP18H04107 and Microsoft Research Asia.

[1] All the words in **C** are from **B**.

5. REFERENCES

[1] Bernstein, J. 2003. Objective measurement of intelligibility. *Proc. ICPhS* 1581–1584.

[2] Derwing, T. M., Munro, M. J. 2015. *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research*. John Benjamins Publishing.

[3] Goslin, J., Duffy, H., Floccia, C. 2012. An erp investigation of regional and foreign accent processing. *Brain and Language* 122(2), 92–102.

[4] Govender, A., King, S. 2018. Using pupillometry to measure the cognitive load of synthetic speech. *Proc. INTERSPEECH* 2838–2842.

[5] Hahne, A. 2001. What's different in second-language processing? evidence from event-related brain potential. *Journal of Psycholinguistic Research* 30(3), 251–266.

[6] Hu, W., Qian, Y., Soong, F. K., Wang, Y. 2015. Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers. *Speech Communication* 67, 154–166.

[7] Inoue, Y., Kabashima, S., Saito, D., Minematsu, N., Kanamura, K., Yamauchi, Y. 2018. A study of objective measurement of comprehensibility through native speakers shadowing of learners' utterances. *Proc. INTERSPEECH* 1651–1655.

[8] Kabashima, S., Inoue, Y., Saito, D., Minematsu, N. 2018. Dnn-based scoring of language learners' proficiency using learners' shadowings and native listeners' responsive shadowings. *Proc. Spoken Language Technology* 971–978.

[9] Kurata, K. 2007. A fundamental study on the cognitive mechanism of shadowing in japanese – from the view point of starting point of oral reproduction, memory span and sentence structure –. *Bulletin of Graduate School of Education, Hiroshima University* 259–265.

[10] Maekawa, K., Koiso, H., Furui, S., Isahara, H. 2000. Spontaneous speech corpus of japanese. *Proc. LREC* 947–952.

[11] Makino, T., Aoki, R. 2012. English read by japanese phonetic corpus: an interim report. *Research in Langauge* 10(1), 79–95.

[12] Minematsu, N., Okabe, K., Ogaki, K., Hirose, K. 2011. Measurement of objective intelligibility of japanese accented english using erj database. *Proc. INTERSPEECH* 1481–1484.

[13] Munro, M. J., Derwing, T. M. 1995. Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning* 45(1), 73–97.

[14] Munro, M. J., Derwing, T. M. 2006. The functional load principle in esl pronunciation instruction: An exploratory study. *System* 34, 520–531.

[15] Nye, P. W., A., F. C. 2003. Shadowing latency and imitation: The effect of familiarity with the phonetic patterning of english. *Journal of Phonetics* 63–79.

[16] Povey, D., Ghoshal, A., Boulianne, G., Glembek, L. B., Goel, N., Hannemann, M., Motlíček, P., Y., Q., P., S., Silovský, J., Stemmer, G., Veselý, K. 2011. The kaldi speech recognition toolkit. *Proc. ASRU*.

[17] RongNa, A. N., Mori, K., N., S. 2015. Latency analysis of speech shadowing reveals processing differences in japanese adults who do and do not stutter. *Proc. INTERSPEECH* 2972–2976.

[18] Song, J., Iverson, P. 2018. Listening effort during speech perception enhances auditory and lexical processing for non-native listeners and accents. *Cognition* 179, 163–170.

[19] Swan, M., Smith, B. 2001. *Learner English –A teacher's guide to interference and other problems–* . Cambridge University Press.

[20] Witt, S. M., Young, S. J. 2001. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication* 30(1), 95–108.

[21] Yue, J., Shiozawa, F., Toyama, S., Yamauchi, Y., Ito, K., Saito, D., Minematsu, N. 2017. Automatic scoring of shadowing speech based on dnn posteriors and their dtw. *Proc. INTERSPEECH* 1422–1426.