A Study of Objective Measurement of Comprehensibility Through Native Speakers' Shadowing of Learners' Utterances

Yusuke INOUE[1], *Suguru KABASHIMA*[1], *Daisuke SAITO*[1], *Nobuaki MINEMATSU*[1]
Kumi KANAMURA[2], *Yutaka YAMAUCHI*[3]

[1]The University of Tokyo, [2]Nagoya University of Economics, [3]Tokyo International University

{inoue0124,kabashima,dsk_saito,mine}@gavo.t.u-tokyo.ac.jp
kanamura@nagoya-ku.ac.jp, yyama@tiu.ac.jp

Abstract

While learners desire to acquire so comprehensible pronunciations as to make themselves understood smoothly, acquisition often becomes difficult because, outside of classrooms, it is not rare that learners can hardly find chances to talk in the target language. Even when they talk to native speakers, they may receive only lenient or superficial suggestions from native speakers. How can learners know native speakers' honest perception on their utterances? In this paper, shadowing is introduced not to learners but to native listeners, who are asked to shadow learners' utterances. Since shadowing is as simultaneous repetition as possible, it is expected that native listeners' perceived comprehensibility can be measured objectively as smoothness of natives' shadowings. Experiments show that 1) shadowers' subjective assessment of learners' speech and that of their shadowings are highly correlated and that 2) the former is more correlated with the GOP scores of natives' shadowings than those of learners' speech. These results suggest it is valid to regard comprehensible pronunciation as *shadowable* pronunciation.
Index Terms: language learning, objective measurement, comprehensibility, natives' shadowing, GOP

1. Background and objective

To become a good user of a new language, a learner has to acquire good skills of speaking, listening, writing, and reading. Among these, acquisition of speaking and listening skills requires oral interactions with others. Further, since listening skills can be improved with classical audio materials, speaking skills may require oral interactions the most, where learners can learn what kind of mispronunciations are more fatal.

To provide those situations of oral interaction technically to learners, dialogue-based CALL (Computer Aided Language Learning) systems have been developed [1, 2, 3], where not only pronunciation errors but also grammatical errors can be detected and their corrective feedback is also provided. To assess learners' pronunciation, native speakers' acoustic models are often referred to and comparison is made between learners' speech and its corresponding native model. Inadequate phonetic realization of phonemes due to foreign accents are automatically detected, but it is a well-known fact that some types of foreign accents hardly reduce smoothness of communication [4, 5, 6].

As far as the authors are aware, users of English accept a large variety of accented pronunciations because English is the primary language used for international communication. Further, probably due to political reasons, English has been accepted as an official language in many countries such as India, Singapore, Philippines, etc, where native people speak English with their own accents and they often recognize their accents as racial identity. The term of World Englishes [7, 8] characterizes well the current state of the language of English.

However, it is still a fact that some foreign accented pronunciations even of English still cause miscommunications and learners want to know what kind of mispronunciations are critical. In short, most of the learners desire to acquire intelligible enough or comprehensible enough pronunciations. In applied linguistics, intelligibility and comprehensibility are defined somewhat differently as follows [4]. 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 by asking native speakers to write down that utterance word by word. Correct identification rate can represent the intelligibility of that utterance.

Comprehensibility of an utterance means how easily and smoothly listeners can understand the content of that utterance, often quantified using questionnaires imposed on listeners. Since correct comprehension of an utterance often requires correct identification of words, the authors consider that comprehensibility covers intelligibility and represents more. Even if all the words of an utterance can be identified correctly but some listening efforts are required for comprehension, that utterance is not treated as highly comprehensible. Previous works [4, 5, 6] showed that some foreign accents can hardly reduce intelligibility and even comprehensibility. Practically speaking, correction of those accented pronunciations may not be needed and corrections are needed primarily for fatal mispronunciations.

How can learners know which parts of their utterances prevent smooth comprehension. [5] shows a possible answer using the notion of functional load based on the impact of phoneme substitution. The authors take another approach, which observes native listeners' behaviors and predicts their perceived comprehensibility. When learners have chances to talk to native speakers, however, they may give only lenient and superficial comments on the learners' pronunciation. Lenient suggestions are good to beginning learners for encouragement, but honest suggestions are often requested from advanced learners.

In this paper, to disclose native listeners' honest perception on learners' utterances and hopefully to investigate which parts of them prevent smooth communication, shadowing learners' utterances is imposed on native listeners. Since shadowing is as simultaneous repetition as possible, native listeners' perceived comprehensibility will be characterized intactly and measured objectively as smoothness of natives' shadowings. As far as the authors know, this is the first attempt in L2 studies to introduce shadowing entirely to native speakers and is also the first attempt to measure comprehensibility objectively. To this end, DNN-based GOP (Goodness Of Pronunciation) scores are calculated both from learners' utterances and natives' shadowings. Then, these objective scores are compared to subjective scores rated on learners' utterances and natives' shadowings. Results show that natives' shadowings are more adequate observations than learners' speech for automatic rating of comprehensibility.

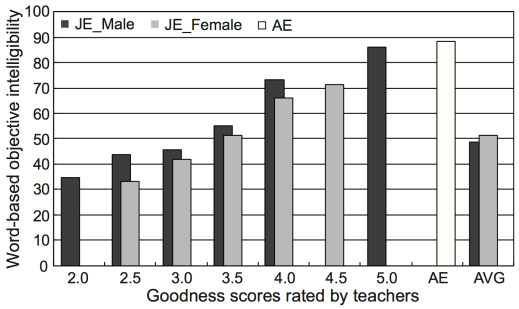

Figure 1: *Word-based intelligibility for seven learner groups [10]*

2. Objective measurement of intelligibility

Objective measurement of intelligibility was made in [9, 10], where English spoken by immigrants to the USA [9] and by Japanese college students [10] were presented to American English native listeners on a telephone line. They were asked, after listening, not to write down but to repeat what they heard. Their repetitions were transcribed word by word manually by technical staff to derive word-based intelligibility of each utterance.

Intelligibility of an utterance is assumed to depend on listeners' prior exposure to speakers' accents. In [10], only American English native speakers who had almost no experiences of talking with Japanese people were adopted as listeners. 800 Japanese-English utterances, collected from 200 Japanese college students, four utterances per student, were presented to 173 listeners. On average, one utterance was presented to 20 listeners. Figure 1 shows word-based objective intelligibility scores of each student group, where 200 students were divided into seven groups based on their overall proficiency scores rated subjectively by teachers. The averaged word-based intelligibility score, i.e. averaged identification rate, is so low as around 50%, while the score of native utterances is approximately 90%[1].

Figure 1 was obtained by observing listeners' behaviors of repetition but no good control was made on listeners' repetition. In this case, it is highly speculated that efforts of listening and delay of repetition depended on listeners. If delay is minimized, repetition becomes shadowing, where only small listening efforts are allowed. The authors consider that results of *repetition* indicate how *intelligible* a given utterance is and that results of *shadowing* indicate how *comprehensible* it is. Further, when repetition is asked to do, a long utterance should not be used for experiments because, in that case, some words in the beginning of an utterance will be forgotten when listeners start repetition. In the case of shadowing, however, long utterances cause no problem. In [10], only native listeners with almost no exposure to Japanese English were adopted. In this paper, different levels of exposure are considered for preparing native listeners.

3. Smoothness of shadowings

To quantify smoothness of shadowings, two speech features are focused on. One is related to accuracy of articulation and the other is to delay of shadowing. For the former, GOP [11] is adopted because it is widely used as baseline feature to indicate accuracy of articulation and, in our previous studies, it was applied to assess learners' shadowing performance [12, 13, 14].

[1]The experimental condition was somewhat artificial. Speaker identity of input utterances was changed sentence by sentence, and the content of any two contiguous sentences was always independent.

Figure 2: *Karaoke-style reading aloud and its recording*

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 [11, 12, 13], classical and generative speech models of HMM (Hidden Markov Models) were used to calculate GOP and, these days, recent and discriminative speech models of DNN (Deep Neural Network) were introduced [14, 15]. DNN experimentally gave a better performance in the task of error detection or proficiency prediction. In this paper, DNN-based GOP scores were calculated both for learners' speeches and natives' shadowings. After forced alignment, the phoneme intended at time t, p_t, was obtained. Then, $P(p_t|o_t)$ was accumulated during an entire utterance. Then, the GOP score of a given utterance x is calculated as follows.

$$\mathrm{GOP}(x) = \frac{1}{D_x} \sum_t P(p_t|o_t), \tag{1}$$

where D_x is the frame-based duration of that utterance [14].

As for delay of shadowing, by comparing forced alignment of a learner's speech and that of its corresponding native 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 utterances were averaged to define delay of shadowing between the two utterances. Generally speaking, shadowing is performed with delay of 1 to 2 seconds to a presented utterance.

4. Experiments of natives' shadowing

4.1. Speaking rate control for speech collection

In this study, the target language of learning is set to Japanese, and learners are Vietnamese. For experiments, their Japanese utterances were collected and, as reference, native Japanese utterances were also collected. In the following section, native listeners are asked to shadow these two types of utterances. If these utterances are very slow, their comprehensibility may be always high and independent of how heavily accented they are.

This is why speaking rate control was introduced for speech collection. At first, an intermediate-level textbook of Japanese with an audio CD was selected [16]. From the CD, six read-aloud paragraphs were adopted. Here, the paragraphs including proper nouns were excluded. A tool of calculating readability, Jreadability [17], verified that the six paragraphs belong to the same readability level. In addition to the professional model speaker's utterances in the CD, each phrase in the six paragraphs was read aloud by six Vietnamese learners (three males and three females) and six native speakers (three males and three females) with their speaking rate being contolled by using a *Karaoka*-style recording program, shown in Figure 2. Forced alignment was performed on the model speaker's utterances, and each written phrase was shown, where the color of text changed according to the model speaker's speaking rate. If a reader stammers, s/he was allowed to read as many times as s/he wanted. Among the six Vietnamese learners, three are at an intermediate level, whose length of learning is shorter than three years (2.7 years on average) and the other three are at

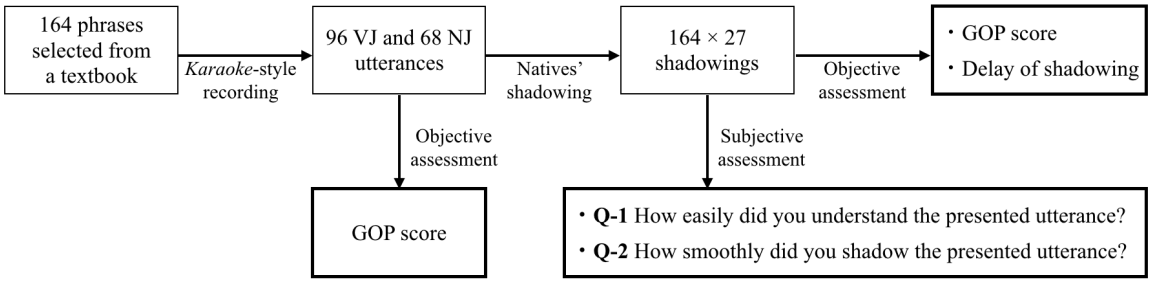

Figure 3: *Overview of the experiments of natives' shadowing*

Table 1: *Averaged scores of comprehensibility and smoothness for Vietnamese Japanese (VJ)*

	NS-1	NS-2	NS-3
comprehensibility (S_C)	4.13	4.20	4.24
smoothness (S_S)	4.58	4.45	4.78

an advanced level, who had learned Japanese longer than three years (5.8 years on average). Finally, 96 Vietnamese Japanese (VJ) utterances and 68 native Japanese (NJ) utterances were collected. The sentences of the two sets are not shared.

4.2. Shadowers' prior exposure to Vietnamese Japanese

As easily expected, easiness of shadowing VJ strongly depends on shadowers' prior exposure to VJ. In this study, we prepared three groups of native speakers.

NS-1 Those who never talked with Vietnamese people.

NS-2 University students whose laboratory has a Vietnamese student who can speak in Japanese.

NS-3 Teachers of Japanese, who have expert knowledge of VJ.

Seventeen, five, and five Japanese adults participated in the experiments as subjects of **NS-1**, **NS-2**, and **NS-3**, respectively. They were instructed to shadow presented utterances, but not to imitate accented pronunciations in the utterances because subjects of **NS-3** can behave like Vietnamese learners when shadowing. They were instructed to shadow a given utterance in native Japanese. Before the experiments, 15-min shadowing practices were made with utterances not used in the experiments.

4.3. Subjective and objective assessment of shadowing

The 27 Japanese participated in the experiment. Each of them shadowed 200 utterances, which are 96 VJ, 68 NJ, and 36 dummy VJ utterances selected from the corpus of Japanese Read by Foreigners (JRF) [18]. Presentation was done only once in random order through headphones and recording was done using a uni-directional ear-hook microphone.

After shadowing, two questions were always asked.

Q-1 How easily did you understand the presented utterance?

Q-2 How smoothly did you shadow the presented utterance?

The former is related to comprehensibility of the presented utterance and the latter is to smoothness of shadowing. In both questions, a shadower uses a seven-degree scale for answering, where higher scores mean easier or smoother. The two scores are expected to be highly correlated in nature, but if strategic differences of rating are found between the two measures in some shadowers, the correlation will be low for them.

Objective assessment was also made to presented utterances (learners' readings) as well as natives' shadowings. The DNN-based GOP score, S_G in short, was calculated for each of learners' readings and natives' shadowings. Delay of shadowing was also calculated for each shadowing. The overview of the experiments of natives' shadowing is illustrated in Figure 3.

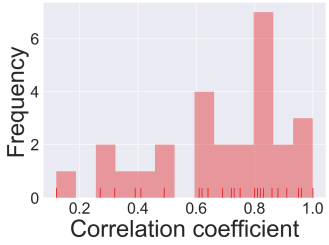

Figure 4: *The histogram of the shadowers' correlations*

5. Results and discussion

5.1. Shadower-dependent subjective rating

Shadowers with different exposures are expected to give different comprehensibility scores, S_C, and different smoothness scores, S_S, to one and the same utterance. Table 1 shows the averaged scores of the two measures for the three groups of shadowers, calculated only from the VJ utterances. ANOVA shows significant differences ($p<0.05$) only between **NS-1** and **NS-3** and between **NS-2** and **NS-3** in the scores of S_S. Although significant differences are not found between any two cases in S_C, a trend of increase can be seen from **NS-1** to **NS-3** in S_C.

5.2. Correlation between the two measures

Correlation between the two kinds of scores, S_C and S_S, is calculated for each subject. Their average is 0.68, which is not so high as expected. Figure 4 shows the histogram of the shadowers' correlations. Seven shadowers out of 27 show very low correlations and their average is 0.36. This is probably because of inter-measure strategic differences exhibited by the seven shadowers. To reduce these differences, some prior discussion should have been done to achieve a consensus on what scores should be given to what kind of learners' readings and to what kind of natives' shadowings. Even with a deep discussion for consensus, however, some strategic differences are inevitable. The averaged correlation among the remaining 20 shadowers is so high as 0.79. In the following section, the authors discuss the results of calculating S_G and delay of shadowing as objective observations. In addition to the results from all the 27 shadowers, those from the above 20 shadowers are also described.

5.3. DNN-GOP and the two subjective scores

Each of the 96 VJ utterances and the 68 NJ utterances had 27 shadowings, 27 S_Cs, and 27 S_Ss. For each shadowing, its S_G

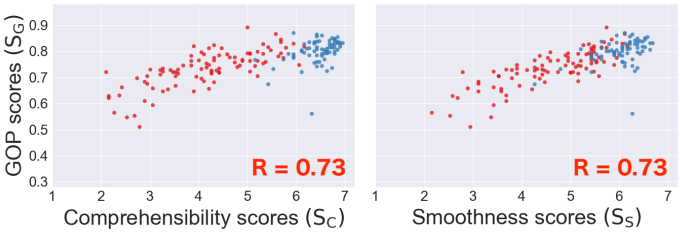

Figure 5: *GOP of shadowings and the two subjective scores*

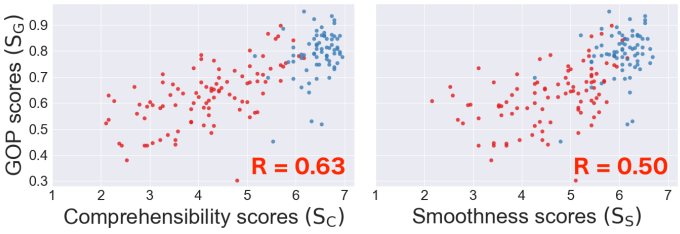

Figure 6: *GOP of learners' speech and the two subjective scores*

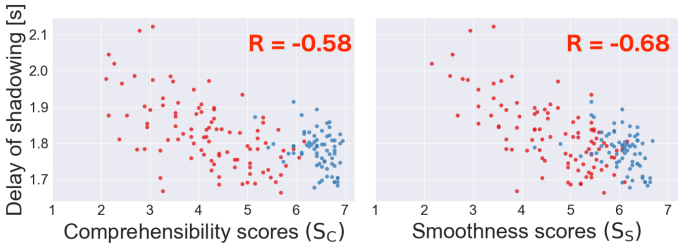

Figure 7: *Delay of shadowing and the two subjective scores*

was calculated. Then, for each of the 164 utterances, the averaged S_G, the averaged S_C, the averaged S_S was calculated over the shadowers. Figure 5 shows the correlation between the averaged S_Cs and the averaged S_Gs, and that between the averaged S_Ss and the averaged S_Gs. Red dots and blue dots are VJ utterances and NJ utterances, respectively. **R** in each figure is the correlation calculated only from the VJ utterances. Both S_G and S_S are scores obtained directly from shadowings and it is natural that they are highly correlated. It is interesting that, although S_C is a score not for shadowings but for learners' readings, S_G and S_C are so highly correlated as S_G and S_S are. When calculating those correlations only from the 20 subjects selected in Section 5.2, **R**(S_G, S_C)=0.75 and **R**(S_G, S_S)=0.72. These results clearly show the validity of using S_G to indicate how comprehensible an input utterance is.

S_Gs are also calculated from the VJ utterances, not from natives' shadowings. The correlations between learners' S_Gs and the two subjective scores are shown in Figure 6. The correlations are definitely lower than those in Figure 5, although S_Gs in this figure and S_Cs are obtained directly from learners' readings. The authors can emphasize that, for GOP-based comprehensibility assessment of learners' utterances, natives' shadowings are much more adequate than learners' utterances to analyze acoustically. This is probably because of two reasons. Comprehensibility is listener-dependent in nature, but S_G calculated from learners' speech is totally independent of listeners. Further, calculation of S_G from native shadowings may be technically stable compared to that from non-native readings.

5.4. Delay of shadowing and the two subjective scores

Figure 7 shows the correlations between delay of shadowing and the two subjective scores of S_C and S_S. As is expected, negative correlations are found in the two figures. The absolute value of the correlation is higher in the case of S_S and the authors consider that this is quite natural because delayed shadowing decreases S_S immediately but not always decreases S_C.

In this section, S_G of natives' shadowings, S_G of learners' readings, and delay of shadowing are analyzed separately but they can be combined in a regression model to automatically predict shadowers' comprehensibility in a more accurate way. This will be done soon as one of the future works.

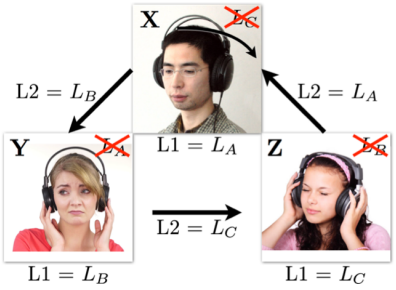

Figure 8: *Inter-learner shadowing*

5.5. Toward inter-learner shadowing

In this paper, a novel and promising framework of objective and automatic measurement of perceived comprehensibility was proposed based on native listeners' shadowing. However, one critical issue has not been discussed so far. How can a sufficient number of native listeners be prepared for learners? The authors are optimistic about this issue because inter-learner shadowing will solve it. Any learner is a native speaker of a language, who is qualified sufficiently to shadow utterances of other learners who are learning that language. Figure 8 explains inter-learner shadowing. Learner **X**, who speaks L_A as L1 and is learning L_B but does not speak L_C, reads aloud sentences in L_B. His utterances in L_B are shadowed by learner **Y**, who speaks L_B as L1 and is learning L_C but does not speak L_A. Her utterances in L_C are shadowed by learner **Z**, who speaks L_C as L1 and is learning L_A but does not speak L_B. Her utterances in L_A are shadowed by learner **X**. The authors consider that this is a speech version of *Lang-8* [19], where any learner can support other learners and can be supported by other learners. If this infrastructure is realized and provided for learners, the above issue will be solved and any learner can be exposed easily to native listeners' honest perception on his/her utterances.

6. Conclusions

In this paper, objective measurement of comprehensibility of learners' readings was examined based on acoustic analysis and GOP calculation of native listeners' shadowings. Experiments showed remarkably promising results. Preparation of native shadowers will be made possible by mutual assistance among learners, i.e. inter-learner shadowing. This framework will solve the problem of lack of exposure partly. As future work, the authors are going to test the proposed method using three groups of learners (L_A, L_B, L_C) = (Japanese, American English, Chinese). Further, the speech samples used for intelligibility measurement [10] will be used for comprehensibility measurement by using native shadowers.

This work was financially supported by JSPS or MEXT KAKENHI JP26118002 and JP26240022.

7. References

[1] Reima Karhila, Sari Ylinen, Seppo Enarvi, Kalle Palomäki, Aleksander Nikulin Olli Rantula, Vertti Viitanen, Krupakar Dhinakaran, Anna-Riikka Smolander, Heini Kallio, Katja Junttila, Maria Uther, Perttu Hämäläinen, and Mikko Kurimo, "SIAK-agame for foreign language pronunciation learning," *Proc. INTERSPEECH*, 3429–3430, 2017.

[2] Wei Li, Sabato Marco Siniscalchi, Nancy F. Chen, and Chin-Hui Lee, "Improving non-native mispronunciation detection and enriching diagnostic feedback with DNN-based speech attribute modeling," *Proc. ICASSP*, 6135–6139, 2016.

[3] Wei Li, Kehuang Li, Sabato Marco Siniscalchi, Nancy F. Chen, and Chin-Hui Lee, "Detecting mispronunciations of L2 learners and providing corrective feedback using knowledge-guided and data-driven decision trees," *Proc. INTERSPEECH*, 3127–3131, 2016.

[4] Murray J. Munro and Tracey M. Derwing, "Foreign accent, comprehensibility, and intelligibility in the speech of second language learners," *Language Learning*, 45, 1, 73–97, 1995.

[5] Murray J. Munro and Tracey M. Derwing, "The functional load principle in ESL pronunciation instruction: An exploratory study, " *System* 34, 520–531, 2006.

[6] Tracey M. Derwing and Murray J. Munro, *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research,* published by John Benjamins Publishing, 2015.

[7] Braj B. Kachru, Yamuna Kachru, and Cecil L. Nelson, *The handbook of World Englishes*, published by Wiley-Blackwell, 2009.

[8] Jennifer Jenkins, *World Englishes: a resource book for students,* published by Routledge, 2009.

[9] Jared Bernstein, "Objective measurement of intelligibility," *Proc. ICPhS*, 1581–1584, 2003.

[10] Nobuaki Minematsu, Kohji Okabe, Keisuke Ogaki, and Keikichi Hirose, "Measurement of objective intelligibility of Japanese accented English using ERJ database," *Proc. INTERSPEECH*, 1481–1484, 2011.

[11] Silke M. Witt and Steve J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, 30, 1, 95–108, 2000.

[12] Luo Dean, Nobuaki Minematsu, Yutaka Yamauchi, and Keikichi Hirose, "Automatic assessment of language proficiency through shadowing," *Proc. ISCLP*, 1–4, 2008.

[13] Luo Dean, Nobuaki Minematsu, Yutaka Yamauchi, and Keikichi Hirose, "Analysis and comparison of automatic language proficiency assessment between shadowed sentences and read sentences," *Proc. SLaTE*, 37–40, 2009.

[14] Junwei Yue, Fumiya Shiozawa, Shohei Toyama, Yutaka Yamauchi, Kayoko Ito, Daisuke Saito, and Nobuaki Minematsu, " Automatic scoring of shadowing speech based on DNN posteriors and their DTW," *Proc. INTERSPEECH*, 1422–1426, 2017.

[15] Wenping Hu, Yao Qian, and Frank K. Soong, "An improved DNN-based approach to mispronunciation detection and diagnosis of L2 learners' speech," *Proc. SLaTE*, 71–76, 2015.

[16] Mariko Matsuura, Shunsui Fukuchi, Maiko Kohno, Kayo Yoshida, *Japanese speech training,* published by ASK publisher, 2014.

[17] Jreadability, https://jreadability.net

[18] Kikuko Nishina, Yumiko Yoshimura, Izumi Saita, Yoko Takai, Kikuo Maekawa, Nobuaki Minematsu, Seiichi Nakagawa, Shozo Makino, Masatake Dantsuji, "Speech database construction for Japanese as second language learning," *Proc. O-COCOSDA*, 187–192, 2002.

[19] *Lang-8*, http://lang-8.com