Interspeech 2022
18-22 September 2022, Incheon, Korea

Detection of Learners' Listening Breakdown with Oral Dictation and Its Use to Model Listening Skill Improvement Exclusively Through Shadowing

Takuya Kunihara[1], *Chuanbo Zhu*[1], *Daisuke Saito*[1], *Nobuaki Minematsu*[1], *Noriko Nakanishi*[2]

[1]Graduate School of Engineering, The University of Tokyo
[2]Faculty of Global Communication, Kobe Gakuin University

{kunihara,zhuchb,dsk_saito,mine}@gavo.t.u-tokyo.ac.jp, nakanisi@gc.kobegakuin.ac.jp

Abstract

In language learners' speech, mispronounced words, word fragments, repairs, filled pauses, etc are often found, and they can be detected with ASR-based CALL systems. When learners are listening, some segments in a given utterance are often difficult to identify or misidentified due to lack of listening skill. In this study, we aim at detecting learners' listening breakdown to measure their listening skill. Listening skill is often quantified by imposing manual dictation on learners, but it has inevitable problems because manual dictation is generally an offline task. To solve the problems, oral dictation is imposed instead, and speaking breakdown is detected in the dictation utterances. Here, we assume that learners' speaking breakdown is attributed to their listening breakdown. This method is applied to measure their listening skill and to model its improvement exclusively through shadowing, which is oral dictation with a short delay and was introduced to language education originally as listening training. 35 Japanese university students attended a 42-day intensive shadowing training, and their shadowing utterances were analyzed to detect listening breakdown. Our model exhibits very monotonous improvement of listening skill as a function of how many days learners attended shadowing.

Index Terms: language learning, listening breakdown, oral dictation and shadowing, PPG-DTW, logistic regression

1. Introduction

Application of speech and language technologies to supporting language learners has been discussed in a large number of research articles for some decades [1–3], and these days, various types of CALL (Computer-Aided Language Learning) applications are found even on smart phones. Some software can calculate a holistic score with respect to specific aspects of learners' speech, such as fluency, intelligibility, comprehensibility, and accentedness [4–10], while others can detect pronunciation errors, filled pauses, word fragments, repairs, etc found in learners' speech [11–13]. Speech technologies can support learners effectively to learn how to speak in a new language.

In this paper, also with speech technologies, we attempt to support learners to learn how to listen in a new language, although listening behaviors cannot be observed acoustically because listening generates no acoustic signals by itself. Due to insufficient skill of listening, for a given utterance, some segments are difficult for learners to identify, thus they may be misidentified. Especially, Japanese learners of English have unique difficulty in listening. The fundamental unit of speech production in Japanese is mora, which is CV or V [14], hence every word boundary has V as preceding segment and C or V as succeeding one. In English, however, a word boundary can have any kind of segments as preceding and succeeding. When both preceding and succeeding segments are consonants, Japanese learners may have troubles in word segmentation. Further, at a word boundary, its preceding and succeeding segments often change phonetically in English, such as linking, elision, assimilation, weakening, etc. Because of these phenomena, learners have listening troubles, or listening breakdown. How to detect their listening breakdown with speech technologies?

Learners' listening skill is often measured by having learners dictate a given utterance [15], which is generally read speech by a native speaker with a script. Word-based dictation accuracy, calculated with the script for reading aloud and the result of manual dictation, can quantify the listening skill. This method, however, has inevitable problems [16]. Since manual dictation has to be made after one-time presentation of the model utterance, it has to be short enough. As manual dictation takes a longer time than the duration of the given utterance, learners can rephrase what they actually heard. These problems are inevitable because manual dictation is generally an offline task. How to observe learners' listening breakdown online?

To measure the speaking skill of a learner, manual dictation is often conducted not by the learner but by evaluators to rate the learner's speaking skill [17]. Here, correct dictation rate is generally referred to as intelligibility. If the learner is understood by the evaluators without any listening breakdown, his/her utterances are rated as intelligible. Even when it is evaluators, not learners, who dictate given utterances manually, the above offline problems exist. In applied linguistics, a method of online observation of listening breakdown had been looked for [16].

In our previous work, a method of online observation of evaluators' listening breakdown was proposed [18], which was developed with speech technologies. The task of manual dictation was replaced by oral dictation, where speaking breakdown in the oral dictation utterances was automatically detected. If we ask evaluators to conduct oral dictation with as short delay as possible, the task is called shadowing [19, 20]. Generally speaking, shadowing is an online task as the delay is about 1 sec. Further, evaluators can shadow an utterance that is so long as 1 min. The two problems above were solved completely.

This paper applies this method to learners for the first time to observe their listening breakdown. Taking into account that shadowing had been originally introduced to language education as a method of listening training [19], we examine objectively the effect of L2 shadowing practices to reduce learners' listening breakdown and to model gradual improvement of their listening skill exclusively through continuing shadowing practices on a daily basis. As far as the authors know, this paper is the first attempt to apply speech technologies to assess learners' listening skill based on online observation of their listening behaviors. Online observation of listeners was examined using Electroencephalography (EEG) and pupillometry [21–24], but they are too expensive to be used in classrooms. Further, these methods are just for observation, not for enhancing listening skills. Our method requires a PC and a headset only for observation, and can enhance listening skills at the same time.

10.21437/Interspeech.2022-440

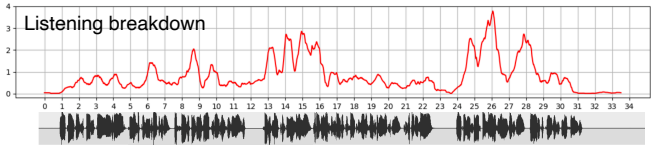

Figure 1: *Listening breakdown quantified via DTW(S,SS)*

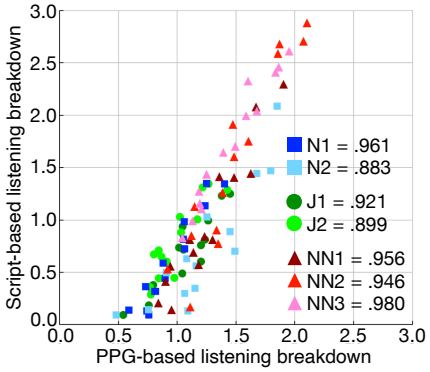

Figure 2: *Corr. between the two kinds of listening breakdown*

2. Related works

2.1. Shadowing-based measurement of sequential listening breakdown while listening to learners' L2 speech [18]

As explained in Section 1, the intelligibility of an L2 utterance was measured objectively and sequentially with shadowing techniques [18]. Seven shadowers, two native speakers and five super-advanced learners, shadowed Japanese-accented English utterances as evaluators. After that, they script-shadowed them. Script-shadowing is a special form of shadowing, where the script of the presented L2 utterance was also shown before and during shadowing. Script-shadowing is functionally equivalent to synchronized reading, and is regarded as shadowing with no listening breakdown. Using shadowing (S) and script-shadowing (SS), Phonetic PosteriorGram-based Dynamic Time Warping was conducted, denoted as PPG-DTW(S,SS), and its alignment path gave us sequential annotation of listening breakdown[1]. One example is shown in Figure 1, from which word-unit listening breakdown was calculated in [18]. By using the script of the L2 utterance and the manual dictation of each evaluator's shadowing, script-based and word-unit listening breakdown was also calculated for each evaluator. Figure 2 shows the correlations between PPG-DTW-based listening breakdown (x-axis) and script-based listening breakdown (y-axis). Blue marks and green marks mean native shadowers and Japanese shadowers with super-advanced English proficiency, respectively. Red marks indicate Chinese or Vietnamese shadowers with super-advanced proficiency, but with no knowledge of Japanese. Irrespective of the evaluators' language backgrounds, we can say that PPG-DTW-based listening breakdown calculation gives us valid and sequential scores of listening breakdown very successfully, even with no manual dictation.

2.2. Shadowing-based enhancement of learners' L2 listening skill [19]

Shadowing was originally proposed in psycholinguistics [25] to analyze human behaviors of perceiving spoken words. In the late 1990s, shadowing was introduced to language education by a Japanese teacher of English [19] as a method of enhancing learners' listening skill. After that, various methods were developed using shadowing techniques for different purposes from listening, such as listening comprehension and pronunciation [20, 26, 27]. Although more than two decades have passed since shadowing was introduced to language education, to the best of the authors' knowledge, automatic assessment of learners' shadowing was always made by comparing shadowing utterances with their model utterances presented for shadowing as in [28]. This is probably because learners were instructed to imitate the model utterances. By revisiting the original purpose of shadowing, which aims at enhancing listening skills, in this paper, we analyze the effect of shadowing on listening skill improvement based on PPG-DTW(S,SS). If readers are interested in why shadowing, i.e. repeating *while* listening, is more effective on listening than listening only or repeating *after* listening, they should refer to the cognitive model proposed in [19, 20].

[1] PPG-DTW(S,SS) directly gives us sequential annotation of the degree of being inarticulate in S compared with SS. Inarticulate productions in S are reasonably attributed to listening breakdown in S.

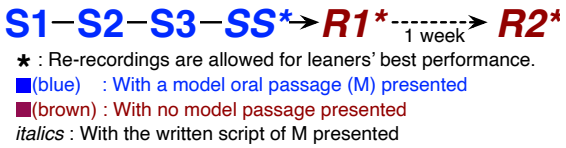

Figure 3: *The task adopted for our shadowing training*

3. Effective reduction of listening breakdown through shadowing practices

3.1. Collection of shadowing and script-shadowing data

We held a very intensive shadowing training for 42 days consecutively, and 35 freshmen or sophomores majoring in Global Communication at a university participated in this training. Their L1 is Japanese. 20 students attended the training every day while the others were absent only on a few days. Before the training, they took Versant Test [10], and their English oral proficiency was assessed to be A1 or A2 on the CEFR scale [29].

In the training, as shown in Figure 3, the participants shadowed a model passage (M) three times (S1, S2, S3) and script-shadowed it once (SS). After SS, they read aloud the script without M (R1). One week later, they read the script again (R2). In this paper, the task of S1-S2-S3-SS-R1-R2 is called as session.

On every day, four sessions were imposed, where four new oral passages were presented. Since shadowing is a task of high cognitive load, the model passages (M) were extracted from the listening part of EIKEN Grade-2 tests [30], which are made for English learners at the level of CEFR A2–B1, and the oral passages were so long as about 30-sec. 168 passages (4 passages $\times$ 42 days) were used in total. For the beginning two weeks, the speaking rate of M reduced by a factor of 0.8. For the next two weeks, it changed by a factor of 0.9 and for the last two weeks, the original passages were used. Speaking rate modification was made with SoX software [31].

It was highly expected that the participants' listening breakdown would depend heavily on the semantic difficulty level of M's content [32], but in the current study, as all Ms were from the EIKEN grade-2 listening tests, the variance of semantic difficulty was considered to be suppressed to some degree. For fair and strict comparison of the participants' listening skill before and after shadowing training, Ms used on Day-01 were used again on Day-23 with about three-week separation, and those on Day-20 were used again on Day-42.

Table 1: *Averaged scores of PPG-DTW(Sn,SS)*

	|S1−SS|	|S2−SS|	|S3−SS|
D-01	1.41	1.28	1.24
D-23	1.09	1.02	0.99
D-20	1.35	1.27	1.24
D-42	1.16	1.10	1.05

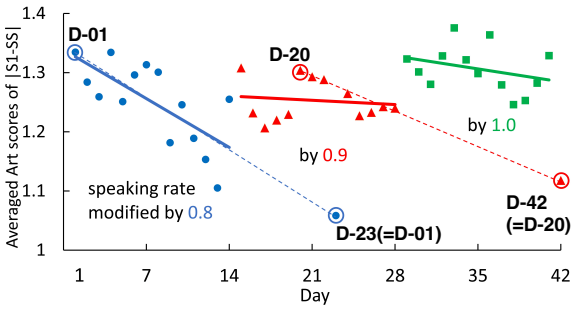

Figure 4: *Gradual reduction of listening breakdown in S1*

In the following section, the averaged PPG-DTW(S,SS) scores are compared between Day-01 and Day-23, and between Day-20 and Day-42. Detailed statistical analysis on the learners' performance of shadowing, i.e. listening and imitation performances, is discussed in another paper [33] with good respect to the prosodic aspect as well as the segmental aspect of speech.

3.2. Effective reduction of listening breakdown

Table 1 shows the averaged scores of PPG-DTW(Sn,SS), denoted simply as |Sn−SS|. PPG extraction was conducted by following [18], where the WSJ-KALDI recipe was used [34]. In each cell, at most 140 PPG-DTW scores (35 learners × 4 sessions) were obtained, and each cell indicates their average. Two-way ANOVA shows that the scores reduce significantly (p<.001) from |S1−SS| to |S3−SS| on every day. By repeating shadowing within a session, listening breakdown reduces effectively. The effect of continuing shadowing practices over days is examined by comparing |Sn−SS| between D-01 and D-23, and between D-20 and D-42. In either case, a significant reduction is found (p<.001) for every n of |Sn−SS|.

Figure 4 shows gradual reduction of the averaged |S1−SS| over 42 days, which was plotted using the data from the 20 participants who attended the training every day. For the first two weeks, the speaking rate of M was modified by 0.8, and for the next two weeks, it changed by 0.9. For the first two weeks, listening skill enhancement is remarkable. For the next two weeks, enhancement became minor, but the scores of D-23 and D-42 are much lower than D-01 and D-20, respectively. As discussed in [19], it was verified again but with no manual operation in this paper that shadowing is effective as listening training. The author of [19] examined his learners' performance by manually dictating the shadowing utterances word by word.

On every day but D-23 and D-42, four different passages were used for shadowing. Although it may be not so large, the level of semantic difficulty may vary from day to day. In the following section, we attempt to remove the semantic difficulty variation to focus on listening breakdown reduction exclusively through continuing shadowing practices over days. For this end, each participant's listening performance after D-day training is modeled with machine learning. By applying the listening performance models with different Ds ($0 \leq D \leq 41$) to the same model passage, we can examine the real effect of continuing shadowing practices over days to improve listening skill.

Table 2: *Various features used to classify an input word segment*

Participant ID p as one-hot vector
Amount of shadowing attended when completing D-day training
Shadowing trial index (n in Sn)
(Lexicosyntactic features)
Part of speech
Dependency parsing
Morphological analysis
Forward position of the word in M
Backword position of the word in M
(Phonological features)
#graphemes
#phonemes
#vowels (=#syllables)
Averaged Phonetic PosteriorGram (PPG)
Holistic speaking rate (0.8, 0.9, or 1.0)
Local speaking rate measured as #vowels per unit time in the word
Preceding and succeeding segments located at the beginning word boundary

Table 3: *The number of word segments in training and testing*

class	θ	training	testing
easy	≤1.5	360,534	115,866
difficult	>1.5	118,650	42,630
sum	—	479,184	158,496

4. Modeling shadowing-based reduction of listening breakdown

4.1. Features used to model each participant's performance

What kind of factors influence PPG-DTW(Sn,SS) in Table 1 and Figure 4? In Table 1, on every day, it is valid that |S1−SS| > |S2−SS| > |S3−SS|. This reduction may be due not to shadowing itself but to memorizing a part of the presented model passage for the 2nd or 3rd shadowing. To focus on listening breakdown reduction achieved exclusively by continuing shadowing practices over days, irrelevant factors should be removed adequately. For this, each participant's performance of listening is modeled and the obtained model is used for analysis.

Let us suppose that the model passage at the i-th session on Day-d, M_i^d, is presented to participant p, who has completed D-day shadowing training so far. S/he gives us a pair of Sn and SS. With the utterance pair, the listening breakdown curve was drawn as in Figure 1[2]. From this curve, the averaged score of PPG-DTW(Sn,SS) for M_i^d was calculated for the j-th word segment, and the score is denoted as $\mathrm{s}_j(\mathrm{M}_i^d,p,D,n)$. By applying an adequate threshold θ to $\mathrm{s}_j(\mathrm{M}_i^d,p,D,n)$, each word segment in M_i^d was classified as easy-to-identify or difficult-to-identity. It is evident that this classification depends on participant ID p, amount of training D, and shadowing trial index n as well as the j-th word's lexicosyntactic features and phonological features. In the following section, we attempt to build a classification model that can classify automatically whether a given word segment is easy or difficult to identify, depending on various factors listed in Table 2. Some of the features were extracted with spaCy [35]. Since all the participants are Japanese, as explained in Section 1, the preceding and succeeding segments at a word boundary are also included in Table 2. Speaking rate was given to our model in two ways, holistic rate and local rate.

[2] D indicates the amount of shadowing attended so far, while d is just an index to the model passage presented on Day-d. In Section 4.3.2, we apply the listening performance models with different Ds ($0 \leq D \leq 41$) to the model passage presented on a specific day of d.

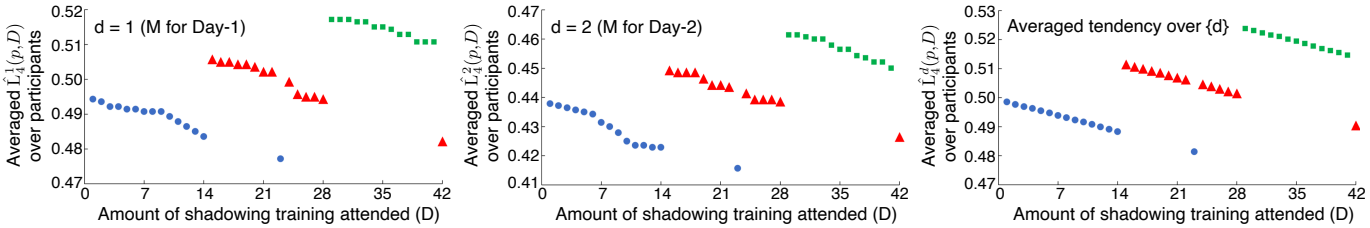

Figure 5: *Simulated reduction of listening breakdown (left: d=1, center: d=2, right: averaged over {d})*

Table 4: *Classification performances*

class	precision	recall	F1-score
Easy	0.857	0.672	0.754
Difficult	0.430	0.688	0.529

Table 5: *Correlations from* $\hat{L}_4^d(p,D)$ *and* R_4^d *to* $L_4^d(p,D)$

$\hat{L}_4^d(p,D)$	FKG	GFI	SMOG	CLI	ARI
0.501	0.090	0.210	0.069	0.016	0.222

4.2. Training data and testing data

All the S and SS utterances of the 20 participants, who attended the shadowing training every day, were used for analysis. The total number of utterances is 10,080 (= 4 sessions per day × 42 days × 3 shadowing trials per session × 20 participants), each of which was so long as about 30 sec. The number of model passages is 168 (= 4 passages per day × 42 days). In the model utterances, the number of word segments is 10,628, and in the participants' responses, it is 637,680. After some preliminary analysis, threshold θ was set to 1.5 and all the available data of 637,680 word segments were divided into training and testing. To train and test our listening performance models, the S+SS utterances obtained from the first three sessions were used for training ($1 \leq i \leq 3$) and those obtained from the last session were used for testing (i=4). Table 3 shows the number of word segments used for analysis. The imbalance problem is found between easy and difficult, but all the data samples were used for analysis as they were. For classification, we used logistic regression by taking the size of training data into account.

4.3. Results and discussion

4.3.1. Classification performance

The trained model is expected to predict whether the j-th word segment in M_4^d is easy or difficult to identify when participant p listens to M_4^d in the n-th shadowing trial after attending D-day shadowing practices. Here, the ground truth was obtained by applying threshold θ to $s_j(M_4^d,p,D,n)$, where $1 \leq d \leq 42$, $1 \leq p \leq 20$, $D=d-1$, and $1 \leq n \leq 3$. Table 4 shows the classification performance of our model. Due to the imbalance problem in Table 3, the performance is worse in the case of difficult word segments. To compare the performance of our model with some references, the following procedure was carried out.

We focused on the first shadowing only, where n=1. By focusing on the ground truth for M_4^d of participant p, namely, whether $s_j(M_4^d,p,d-1,1)$ is higher than θ, the ratio of difficult words over the entire words is calculated, which is interpreted as *listenability* and denoted as $L_4^d(p,D)$, where $D=d-1$. Our model can predict $L_4^d(p,D)$, representing how difficult it is for participant p after D-day training to identify the words in M_4^d correctly. The predicted score is written as $\hat{L}_4^d(p,D)$. For comparison, we use *readability* scores, R_4^d, which indicates how difficult for learners to read the script used for M_4^d. There are several well-known methods to calculate readability scores for any text. Here, we use Flesch-Kincaid Grade Level (FKG), Gunning Fog Index (GFI), Simple Measure of Gobbledygook Index (SMOG), Coleman-Liau Index (CLI), and Automated Readability Index (ARI) [36]. In either case of listenability and readability, the higher the score, the more difficult the passage.

Table 5 shows the correlations from $\hat{L}_4^d(p,D)$ and the five kinds of readability scores to the ground truth of $L_4^d(p,D)$. As R_4^d is totally independent of learners' attributes and many of each word's lexicosyntaxtic or phonological features listed in Table 2, our model is by far superior to the readability models.

4.3.2. Simulation of listening breakdown reduction

By applying our model to predict listenability $L_4^d(p,D)$ for M_4^d, we focus on the direct effect of continuing shadowing practices to reduce listening breakdown, or to improve listening skill. As in Section 4.3.1, only the first shadowings were focused on. For each of M_4^d, where $1 \leq d \leq 42$, our models were applied to calculate 42 kinds of listenability, i.e. $\hat{L}_4^d(p,D)$, where $0 \leq D \leq 41$. Here one passage M was tested with 42 different models. It should be noted that our models do not store input passages in memory, thus any input is always new to our models, which is totally different from behaviors of human participants.

Figure 5 shows the averaged $\hat{L}_4^d(p,D)$ over the 20 participants. The left and the center are results of the model passages presented on Day-01 and Day-02, respectively. For the first two weeks, the holistic speaking rate variable in Table 2 was set to 0.8, and it changed to 0.9 and 1.0 after completing tasks of another two weeks. The tendency of listening breakdown reduction is clearly found in either case, and abrupt reductions on Day-23 and Day-42 are also well simulated. For other days ($3 \leq d \leq 42$), similar tendencies were obtained and the average of 42 reduction tendencies is plotted on the right. It is evident that the amount of shadowing training attended is monotonously related to listening breakdown reduction. Although learners' listening performance is expected to reach a plateau after extensive training, our experimental results claim that learners at A1 and A2 CEFR levels can surely improve their listening skills by attending shadowing training, and that the more they attend shadowing, the more skillful they surely become in listening.

5. Conclusions

This paper is the fist attempt to detect learners' listening breakdown with speech technologies, where oral dictation, i.e. shadowing was imposed on learners for measurement. Taking into account that shadowing is effective for learners to improve their listening skill, our method was used to observe their gradual improvements. By modeling their improvements, the direct effects of shadowing was discussed based on numerical simulation. We're interested in sophisticating our models so that they can be used to characterize weakness of listening of individual learners and to select adequate listening material for them.

6. References

[1] M. Eskenazi, "An overview of spoken language technology for education," *Speech Communication*, vol. 51, pp. 832–844, 2009.

[2] T. Kawahara and N. Minematsu, "Computer-Assisted Language Learning (CALL) based on speech technologies," *IEICE Trans. Info. Sys.*, vol. J96-D, no. 7, pp. 1549–1565, 2013.

[3] T. Isaacs, "Fully automated speaking assessments: changes to proficiency testing and the role of pronunciation," in *The Routledge handbook of contemporary English pronunciation*, O. Kang, R. I. Thomson, and J. Murphy, Eds. Routledge, 2018, pp. 570–584.

[4] E. Ribeiro, J. Ferreira, J. Olcoz, A. Abad, H. Moniz, F. Batista, and I. Trancoso, "Combining multiple approaches to predict the degree of nativeness," in *Proc. INTERSPEECH*, 2015, pp. 488–492.

[5] Y. Xiao, F. Soong, and W. Hu, "Paired phone-posteriors approach to ESL pronunciation quality assessment," in *Proc. INTERSPEECH*, 2018, pp. 1631–1635.

[6] L. Chen, L. Davis, K. Zechner, C. M. Lee, S.-Y. Yoon, M. Ma, K. Evenini, R. Mundkowsky, X. Wang, C. Lu, A. Loukina, C. W. Leong, J. Tao, and B. Gyawali, "Automated scoring of nonnative speech using the SpeechRater v.5.0 engine," *ETS Research Report Series*, vol. RR-18, no. 10, pp. 1–31, 2018.

[7] J. Fu, , Y. Chiba, T. Nose, and A. Ito, "Automatic assessment of English proficiency for Japanese learners without reference sentences based on deep neural network acoustic models," *Speech Communication*, vol. 116, pp. 86–97, 2020.

[8] Y. Shen, A. Yasukagawa, D. Saito, N. Minematsu, and K. Saito, "Optimized prediction of fluency of L2 English based on interpretable network using quantity of phonation and quality of pronunciation," in *Proc. International Workshop of Spoken Language Technology*, 2021, pp. 698–704.

[9] K. Saito, K. Macmillan, M. Kachlicka, T. Kunihara, and N. Minematsu, "Automated assessment of second language comprehensibility: review, training, validation, and generalization studies," *Studies in Second Language Acquisition*, pp. 1–30, 2022.

[10] *Versant English Test*, https://www.pearson.com/english/versant/tests.html.

[11] W. Hu, Y. Qian, and F. K. Soon, "An improved DNN-based approach to mispronunciation detection and diagnosis of L2 learners' speech," in *Proc. SLaTE*, 2015, pp. 71–76.

[12] *ELSA*, https://elsaspeak.com/.

[13] *Liulisho*, https://www.liulishuo.com/.

[14] H. Kubozono, *Handbook of Japanese Phonetics and Phonology*. Mouton De Gruyter, 2015.

[15] H. Jeong, A. Elgemark, and B. Thorén, "Swedish youths as listeners of global Englishes speakers with diverse accents: listener intelligibility, listener comprehensibility, accentedness percetption, and accentedness acceptance," *Frontiers in Education*, vol. 6, p. 206, 2021.

[16] M. J. Munro and T. M. Derwing, "Foreign accent, comprehensibility and intelligibility, redux," *Journal of Second Language Pronunciation*, vol. 6, no. 3, pp. 283–309, 2020.

[17] R. I. Thomson, "Measurement of accentedness, intelligibility, and comprehensibility," in *Assessment in Second Language Pronunciation*, O. Kang and A. Ginther, Eds. Routledge, 2017, pp. 11–29.

[18] C. Zhu, N. Minematsu, and N. Nakanishi, "Multi-granularity annotation of instantaneous intelligibility of learners' utterances based on shadowing techniques," in *Proc. Automatic Speech Recognition and Understanding (ASRU)*, 2021.

[19] K. Tamai, "The effect of shadowing on listening comprehension," in *The Study of current English*, vol. 36, 1997, pp. 105–116.

[20] S. Kadota, *Shadowing as a Practice in Second Language Acquisition: Connecting Inputs and Outputs*. Routledge, 2019.

[21] A. S. Ihara, A. Matsumoto, S. Ojima, J. Katayama, K. Nakamura, Y. Yokota, H. Watanabe, and Y. Naruse, "Prediction of second language proficiency based on electroencephalographic signals measured while listening to natural speech," *Frontiers in Human Neuroscience*, vol. 15, 2021.

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

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

[24] J. Goslin, H. Duffy, and C. Floccia, "An ERP investigation of regional and foreign accent processing," *Brain and Language*, vol. 122, no. 2, pp. 92–102, 2012.

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

[26] S. Miyake, "Cognitive processes in phrase shadowing and EFL," *JACET Bulletin*, vol. 48, pp. 15–28, 2009.

[27] 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.

[28] S. Kabashima, Y. Inoue, D. Saito, and N. Minematsu, "DNN-based scoring of language learners' proficiency using learners' shadowings and native listeners' responsive shadowings," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 971–978.

[29] *Common European Framework of Reference for Languages: Learning, teaching, assessment*, https://www.coe.int/en/web/common-european-framework-reference-languages.

[30] *EIKEN Test*, https://www.eiken.or.jp/eiken/en/.

[31] *SoX, Sound eXchange*, http://sox.sourceforge.net.

[32] T. Trisitichoke, S. Ando, Y. Inoue, D. Saito, and N. Minematsu, "Influence of content variations on smoothness of native speakers' reverse shadowing," in *Proc. ICPhS*, 2019.

[33] T. Kunihara, N. Minematsu, and N. Nakanishi, "Gradual improvements observed in learners' perception and production of L2 sounds through continuing shadowing practices on a daily basis," in *Proc. INTERSPEECH*, 2022 (to be published).

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

[35] *spaCy*, https://spacy.io/.

[36] *Readability Formulas*, https://readabilityformulas.com/free-readability-formula-tests.php