INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Automatic Prediction of Language Learners' Listenability Using Speech and Text Features Extracted from Listening Drills

Yingxiang Gao†, *Jaehyun Choi*†, *Nobuaki Minematsu*†, *Noriko Nakanishi*‡, *Daisuke Saito*†

†The University of Tokyo, Japan, ‡Kobe Gakuin University, Japan

{gyx,jhchoi2178,mine,dsk_saito}@gavo.t.u-tokyo.ac.jp, nakanisi@gc.kobegakuin.ac.jp

Abstract

When language learners are listening to L2 speech, they experience listening disfluencies or breakdowns not rarely. Although listening disfluencies are mental phenomena, previous studies showed that they can be measured acoustically by asking the learners to shadow the L2 speech, where inarticulate productions in shadowing are reasonably attributed to listening disfluencies. In this paper, we model the measured listening disfluencies by BLSTM and attempt to predict which words in new listening drills are difficult to perceive correctly. Taking some studies in psycholinguistics and applied linguistics into account, which revealed what kind of factors influence human perception of spoken words, speech and text features are extracted from listening drills and used for prediction. Experiments show that our model shows a better performance than other models previously proposed and that learners' factors are very effective for prediction because learners are developing through training.

Index Terms: Language education, listenability, prediction, BLSTM, shadowing, psycholinguistics

1. Introduction

Application of speech technologies to supporting language learners has been discussed for decades [1–5], and recently, various CALL (Computer-Aided Language Learning) applications are running even on smart phones. However, many of them aim at speaking training, where speaking errors are automatically detected and speaking scores are provided to learners [6–9].

Very recently, researchers have started to apply speech technologies to listening training [10–14]. If one follows the strategies adopted when applying speech technologies to speaking training, s/he may have to develop a method to detect listening errors or listening disfluencies, which take place mentally in learners' mind. It seems that researchers took two approaches to listening disfluencies: without or with explicit measurement of the disfluencies. In [10, 11], it was assumed simply that low-performance automatic speech recognition (ASR) systems can simulate learners' performance of listening. With weak ASR, recognition errors were detected in advance from listening drills, and the phrases including the ASR errors were presented in actual listening drills with additional information such as script and/or image to facilitate listening. Although [10, 11] showed some pedagogical effects, where learners' performance was improved, it was not clear how well weak ASR systems can simulate learners' listening behaviors. In [15, 16], unignorable discrepancies were reported between human misperceptions and machine misrecognitions of L2 speech.

In the other approach, researchers attempted to measure learners' listening disfluencies with brain monitoring techniques [13] or without them [14]. Since it is not at all feasible that every learner in class wears an expensive brain sensing device, we focus on measurement of learners' listening disfluencies not from their brains directly but from their brains indirectly [14]. In brain sciences, listening to an input speech is said to cause involuntarily neural activations corresponding to movements of articulatory organs required to reproduce the content of the input speech [17–19]. These neural activations are called mirroring, and in language education, voluntary and vocalized mirroring is called shadowing, which was originally introduced to education to improve the listening skill of learners [20]. In shadowing training, a new model speech is presented to learners without any script, who are asked to reproduce the speech as immediately as possible. It is reasonable to regard shadowing as voluntary and vocalized mirroring and mirroring is as involuntary and silent shadowing [21]. When listening disfluencies take place while shadowing, their shadowing productions become inarticulate easily [22], and in [16], the degree of being inarticulate was sequentially quantified along the time axis. Different from [10, 11], this approach objectively measures learners' listening disfluencies, and in [14], the listening disfluencies were tentatively modeled using logistic regression.

Inspired by [14], in the current study, we attempt to tune this prediction model of learners' listenability. In psycholinguistics, researchers revealed what kind of acoustic, prosodic, syntactic, and pragmatic factors in given speech influence listeners' perception [23, 24]. It should be noted that the influential factors may be different between native listeners and non-native listeners. It should also be noted that, for prediction, we extract speech features and text features explicitly from input speech. These days, without explicit feature extraction, inputs or front ends are mapped directly to outputs or back ends. So-called end-to-end (E2E) systems are beneficial because, if a huge amount of data can be prepared, training of E2E systems does not require any domain-specific knowledge. In some applications such as medical and educational applications, however, end-users including medical doctors and language teachers are often eager to know what kind of features observed in inputs are more influential to prediction [25, 26]. In other words, they want to interpret the mechanism of prediction, and they want glass boxes, not black boxes. Taking their interests into good account, in this paper, after extracting speech and text features explicitly, they are used for prediction. In the experiments, we also compare our proposed model with the weak ASR model.

2. Related works

2.1. Measurement of learners' listening disfluencies

By asking non-native listeners to shadow, their listening disfluencies are sequentially quantified as the degree of being inarticulate in shadowing [16]. For this purpose, a special kind of

10.21437/Interspeech.2023-2541

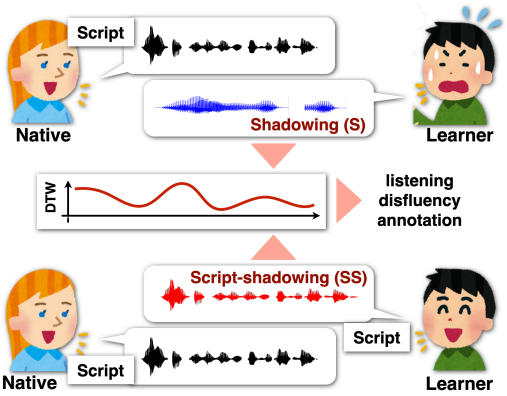


Figure 1: *Measurement of listening disfluencies [16]*

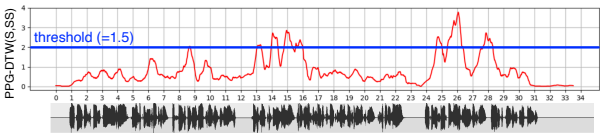


Figure 2: *Sequential annotation of listenability [16]*

shadowing is imposed immediately after shadowing (S), that is script-shadowing (SS), where the script is given to the listeners, shown in Figure 1. In SS, since every word is completely known before shadowing, no listening disfluency happens. Then by comparing S with SS based on Phonetic PosteriorGram-based Dynamic Time Warping (PPG-DTW), the degree of being inarticulate in shadowing, i.e. listening disfluencies, is sequentially quantified. One example is shown as Figure 2. By performing forced alignment on the native speech in Figures 1 and 2, we can calculate the averaged score of a specific learner's listenability for each word in the native speech. By preparing an adequate threshold θ, we can label each word in the native speech as easy-to-listen (PPG-DTW$<\theta$) or difficult-to-listen (PPG-DTW$\geq\theta$) for that specific learner. It is reasonable that listenability of a word in the native speech varies among learners with different learning profiles.

2.2. 42-day shadowing marathon

In [27, 28], an intensive shadowing training was conducted for 42 days running, and 41 university students participated in it. Before the training, they took Versant Test [7], and their L1 was Japanese. On every day, the students participated in four shadowing sessions. In every session, a new and about 30-sec model passage (M) was presented and the students shadowed M three times (S1, S2, and S3). After that, they script-shadowed it once (SS). After SS, they read aloud the script without M (R1). One week later, they read the script aloud again (R2). A session was composed of S1-S2-S3-SS-R1—R2, where repeated recording was allowed only for SS and Rn if some errors were detected.

Since shadowing is a task of high cognitive load, Ms were extracted from the listening part of EIKEN Grade-2 tests [29], which are made for high school students. 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 [30]. For fair comparison between before and after shadowing training, four Ms on Day-01 were used again on Day-23, and those on Day-20 were used again on Day-42. The total number of unique model passages was $4\times42-4\times2=160$. In [28], the effect of shadowing was measured as a declining tendency of the averaged score of PPG-DTW(S1,SS) over the 42 days, shown in Figure 3.

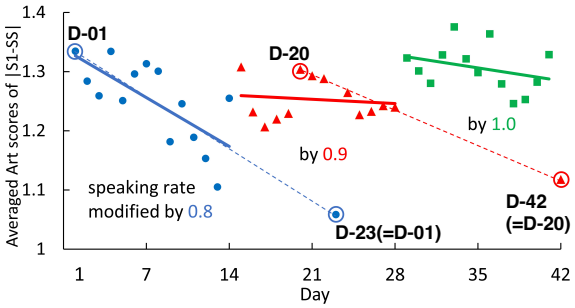


Figure 3: *Gradual reduction of PPG-DTW(S1,SS) [28]*

In [14], the shadowing data collected above were used tentatively to train a model for learners' listenability using logistic regression, which was used to classify each word in new Ms to be easy or difficult to identify. In the current study, using the same corpus, we will test network models for classification and regression. In the latter task, the network model is expected to predict the degree of being inarticulate in shadowing directly for each word in new Ms. Further, these models are compared with the weak ASR models.

2.3. Influential factors to spoken word perception

Psycholinguistically speaking, our task of modeling learners' listenability is regarded as modeling non-native listeners' word perception or misperception. [23, 24] discuss what kind of factors influence general listeners' perception, and [31] discusses learner- and L1-specific factors that cause listening disfluencies. The former explains that general listeners' performance of word perception depends on various factors of the word stimuli such as frequency, age of acquisition, imagery, concreteness, association value, meaningfulness, norm of association, and familiarity. The latter explains that Japanese learners of English have unique difficulty in segmenting an English speech into words. The fundamental unit of speech production in Japanese is mora, which is CV or V [32], 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 ones. When the final segment of a word is C and the initial segment of the following word is V, it is difficult to perceive the word boundary inside the CV. Further, at a word boundary, its preceding and succeeding segments often change phonetically in English, such as linking, elision, assimilation, weakening, etc, which also cause listening troubles frequently to Japanese learners of English.

3. Experiments

3.1. Formulation of the problem to solve

As explained in Section 1, previous studies to support speaking training have proposed various techniques to detect speaking errors from learners' speech. In shadowing training, listening errors, or shadowing errors, can be detected from learners' speech as explained in Section 2.1. In [10, 11], for listening training, the speech segments in new native samples that were expected to cause listening errors were presented with additional hints to facilitate listening. If those segments can be predicted precisely, teachers will select more adequate listening drills differently for individual learners. Taking this application scenario

Table 1: *The number of words in training and testing*

class	θ	training	testing
easy	$\leq$1.5	457,908	147,169
difficult	>1.5	185,514	66,764
sum	—	643,422	213,933

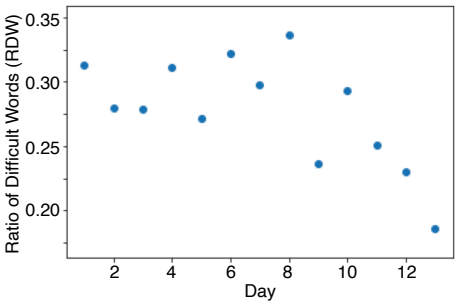


Figure 4: *Gradual reduction of ratio of difficult words (RDW)*

into account, we define the problem to solve in this paper is not detecting but predicting listening errors of individual learners.

To predict how listening disfluency reduces by continuing shadowing practices during the 42-day shadowing marathon, each learner's performance of listening is modeled. Let us suppose that the model passage at the i-th session on Day-d, M_i^d, is presented to learner p, who has completed D-day shadowing training so far[1]. In the shadowing training, Sn (n=1,2,3) and SS were recorded, and the listening disfluency curve was drawn as in Figure 2. From this curve, the averaged score of PPG-DTW(Sn,SS) was calculated for the j-th word in M_i^d, 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 in M_i^d was classified as easy-to-identify or difficult-to-identity. It is evident that this classification depends on learner ID p, amount of training D, and shadowing trial index n as well as the j-th word itself. In the following section, we attempt to build a regression model that can predict $\mathrm{s}_j(\mathrm{M},p,D,n)$ of the j-th word in new M and a classification model that can classify whether the word in M is easy or difficult to identify.

3.2. Training samples and testing samples

In the model utterances, the number of words is 10,628, and in the learners' responses of Sn (n=1,2,3) and SS, it is 857,355. Threshold θ was used to classify each word in the model utterances as easy or difficult depending on the learners' responses. Following [14], θ was set to 1.5. To train and test our listenability models, the Sn+SS utterances obtained from the first three sessions ($1\leq i\leq 3$) were used for training and those obtained from the last session (i=4) were used for testing. It should be noted that word j in M_i^d has multiple scores of $\mathrm{s}_j(\mathrm{M}_i^d,p,D,n)$, depending on p, D, and n. If we take a pair of a word in M and one of its multiple scores as a labeled sample, the numbers of training and testing samples for the two classes are shown in Table 1.

With θ given, it is possible to calculate the ratio of difficult-to-identify words (RDW). On each day for the first two weeks, RDW was calculated by using Sn (n=1,2,3) from all the sessions (i=1,2,3,4) and all the 41 learners, shown in Figure 4. Similar to Figure 3, a clear declining tendency is found. In

[1] D indicates the amount of shadowing attended so far, while d is just an index to the model passage presented on Day-d. On Day-d, the amount of shadowing training attended so far, D, is $d-1$.

Table 2: *Features used for regression and classification* (#xxx indicates the number of xxx.)

Lexicosyntactic features
Part of speech, Dependency parsing, Morphological analysis
Forward position and backward position of the word in M
1-gram and 2-gram scores, JACET8000 scores
Phonological features
#graphemes, #phonemes, #vowels (=#syllables)
Maximum number of consonants consecutively connected
Preceding and succeeding segments at the beginning boundary
Averaged Phonetic PosteriorGram (PPG)
Holistic speaking rate (0.8, 0.9, or 1.0)
Local speaking rate (#vowels per unit time in the word)
Learner and shadowing features
Participant ID p as one-hot vector
Amount of shadowing training attended so far (D, $0\leq D\leq 41$)
Shadowing trial index (n in Sn)

[23, 24, 31], some factors were discussed to influence general listeners' perception and non-native learners' perception, but all the factors discussed are regarded as speech (and text) factors. Figures 3 and 4 indicate that listeners' factors, which should be regarded as developmental factors of listeners, are also influential to their perception. Learners are developing.

3.3. Features used for regression and classification

Table 2 show a full list of features examined in the experiments. We have three kinds of features: 1) lexicosyntactic, 2) phonological, and 3) learner and shadowing features. Some of the lexicosyntactic features were extracted with spaCy [33]. Two scores of n-gram (n=1, 2) and JACET8000 [34] roughly correspond to frequency and age of acquisition, respectively. The latter represents when Japanese learners are supposed to encounter a specific word for the first time on English textbooks. The maximum number of consecutive consonants as well as the preceding and succeeding segments at the beginning boundary are introduced because of Japanese-learner-specific difficulties of segmenting an English speech stream into words.

3.4. Regression models

Bi-directional Long Short-Term Memory (BLSTM) was used to predict $\mathrm{s}_j(\mathrm{M}_4^d,p,D,n)$, where $1\leq d\leq 42$, $1\leq p\leq 41$, $0\leq D\leq 41$, and $1\leq n\leq 4$, where $D=d-1$. BLSTM was trained in two ways: 1) one-step training, where all the features in Table 2 were taken as input to BLSTM, which calculated a predicted score of $\hat{s}_j(\mathrm{M}_4^d,p,D,n)$, and 2) two-step training, where only the lexicosyntactic and phonological features were used for prediction with the first BLSTM, whose output as well as the speaker and shadowing features were used as input to the second BLSTM. The purpose of the two-step training is to reduce the number of parameters of BLSTM, and to examine how effective the learners' factors are for prediction.

In all the experiments for both tasks of regression and classification, BLSTM with time steps of 5 was used, where the features of the current word, its two preceding words, and its two succeeding words were used as input. The number of hidden layers is two with 100 neurons and 50 neurons. To avoid over-fitting, dropout rate was 0.5, and batch size was 256.

3.5. Classification models

Classification was performed in four ways: 1) by following [14], logistic regression was applied, where all the features in Table 2 was used as input, 2) BLSTM was applied to classify

a given word in M_4^d directly, where only the two-step training was used, 3) the two-step BLSTM was applied for regression at first, and then the predicted score of $\hat{s}_j(M_4^d,p,D,n)$ was compared with θ to classify the word j in M_4^d, and 4) recognition results of weak ASR was used for classification, which will be explained in detail in the following paragraph.

3.6. Weak ASR models

ASR systems with higher word error rates were used to simulate learners' listening behaviors. Here, instead of modifying the internal modules or functions in ASR systems, we added white noise to input speech with variable signal-to-noise ratios (S/N [db] $= -20$ to $+20$, with a step of 5) to simulate various word error rates. In our experimental setup, M_i^d (i= 1,2,3) were used for training and M_4^d were used for testing. To simulate learner p on Day-d, RDW was calculated using M_i^d (i= 1,2,3) for learner p on Day-d. The S/N ratio was then selected so that the word error rate of ASR with that S/N ratio can approximate the RDW. After adding white noise to M_4^d with the selected S/N ratio, M_4^d was recognized. The correctly recognized words in M_4^d were judged as easy to learner p on Day-d, and the others were as difficult. Google Chrome ASR was used in the experiments.

By using the four kinds of classification models, we can calculate a predicted score of the ratio of difficult words (RDW) in M_4^d, which is assessed in comparison with its ground truth obtained from $s_j(M_4^d,p,D,n)$ and θ. The prediction performance for RDW is regarded as a holistic metric to assess the classification performance, while accuracy in classifying the individual words is an analytic metric for assessment.

3.7. Readability models

Readability is a text-based measure of how easy or difficult a given text is to read. There are several well-known methods to calculate readability, and the following measures were applied to the text of M_4^d to be compared with the ground truth of RDW. We used Flesch-Kincaid Grade Level (FKG), Gunning Fog Index (GFI), Simple Measure of Gobbledygook Index (SMOG), Coleman-Liau Index (CLI), and Automated Readability Index (ARI) [35]. All the measures are based on superficial properties of text, and the higher the score is, the more difficult the given passage is. Although all the measures ignore the phonological factors and the learner-specific factors listed in Table 2, the above readability scores were calculated for the text of M_4^d.

4. Results

4.1. Performance in prediction

Table 3 shows the performances of prediction. Here, the correlation between the predicted scores, $\hat{s}_j(M_4^d,p,D,n)$, and their ground truth, $s_j(M_4^d,p,D,n)$, is used as metric. Although the difference between one-step BLSTM and two-step BLSTM is minor, the latter shows a slightly higher correlation. The first BLSTM in two-step indicates the prediction performance only with text features and speech features. By comparing this with the second BLSTM in Table 3, we can say that the learner and shadowing features are very influential to prediction.

4.2. Performance in classification

4.2.1. Analytic assessment

As shown in Table 4, the four classification models are compared based on several metrics. If use take accuracy as metric, which represents how many words in M_4^d are correctly classified, BLSTM and BLSTM with thresholding (θ) show better results than the others. As explained in Section 3.1, teachers may want to detect the words in new listening drills which cause listening disfluency. In Table 4, P, R, and F1 are precision, recall, and F1-score of difficult-to-identify words, respectively. The highest F1-score was obtained by BLSTM with thresholding. Our BLSTM models are better than the weak ASR model.

Table 3: *Prediction performances with BLSTM*

BLSTM model	correlation
one-step	0.454
the first BLSTM in two-step	0.355
the second BLSTM in two-step	0.461

Table 4: *Classification performances*

model	accuracy	P	R	F1
logistic regression	0.673	0.457	0.254	0.327
BLSTM	0.734	0.626	0.363	0.460
BLSTM with θ	0.728	0.579	0.472	0.520
weak ASR	0.610	0.347	0.312	0.328

Table 5: *Performance in predicting RDW*

model	correlation
BLSTM with θ	0.694
weak ASR	0.435
FKGL	0.179
GF	0.122
SMOG	-0.003
ARI	0.196
CLI	0.072

4.2.2. Holistic assessment

The ratio of difficult words (RDW) in M_4^d was calculated for each combination of p, D, and n. Then, RDW was predicted by using the classification results of two models of BLSTM with thresholding and weak ASR. The correlations between the ground truth of RDW and the predicted ratios are shown in Table 5, where the correlations between each kind of the readability scores and the ground truth are also presented. Our BLSTM model again shows a higher performance than the week ASR model. It is also clear that the readability models are not useful at all to know how difficult a given listening drill is.

5. Discussion and conclusions

In this paper, the degree of L2 listening disfluency, which was objectively measured through learners' shadowing and script-shadowing, was modeled using networks. The model was applied to predict which words are difficult to identify in new listening drills. Experiments showed that our network models are superior to conventional models in both tasks of prediction and classification. Different from conventional tasks such as speech recognition and speaker recognition, the degree of listening disfluency can vary even when a single speech sample is taken as input. To achieve a good performance, learners' behaviors have to be monitored, and this fact may make it difficult to collect a huge enough amount of data to realize end-to-end prediction and classification. However, shadowing and script-shadowing are widely used in language class in Japan, which may provide researchers with a huge amount of labelled data. By logging individual learners' learning history for months, it will be possible enough to train end-to-end models.

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] 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.

[5] B.-C. Yan, H.-W. Wang, and B. Chen, "PEPPANET: Effective mispronunciation detection and diagnosis leveraging phonetic, phonological, and acoustic cues," in *Proc. Spoken Language Technology*, 2023, pp. 1045–1051.

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

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

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

[10] M. S. Mirzaei, K. Meshgi, and T. Kawahara, "Detecting listening difficulty for second language learners using Automatic Speech Recognition errors," in *Proc. SLaTE*, 2017, pp. 156–160.

[11] M. S. Mirzaei and K. Meshgi, "Adaptive Listening Difficulty Detection for L2 Learners Through Moderating ASR Resources," in *Proc. INTERSPEECH*, 2021, pp. 3929–3933.

[12] H. Zhang, Y. Inoue, D. Saito, N. Minematsu, and Y. Yamauchi, "Computer-aided high variability phonetic training to improve robustness of learners' listening comprehension," in *Proc. ICPhS*, 2019, pp. 924–928.

[13] 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.

[14] T. Kunihara, C. Zhu, D. Saito, N. Minematsu, and N. Nakanishi, "Detection of learners' listening breakdown with oral dictation and its use to model listening skill improvement exclusively through shadowing," in *Proc. INTERSPEECH*, 2022, pp. 4461–4465.

[15] T. M. Derwing, M. J. Munro, and M. Carbonaro, "Does popular speech recognition software work with ESL speech?" *TESOL Quarterly*, vol. 34, no. 3, pp. 592–603, 2000.

[16] C. Zhu, R. Hakoda, D. Saito, N. Minematsu, N. Nakanishi, and T. Nishimura, "Multi-granularity annotation of instantaneous intelligibility of learners' utterances based on shadowing techniques," in *Proc. Automatic Speech Recognition and Understanding*, 2021, pp. 1071–1078.

[17] S. M. Wilson, A. P. Saygin, M. I. Sereno, and M. Iacoboni, "Listening to speech activates motor areas involved in speech production," *Nature Neuroscience*, vol. 7, no. 7, pp. 701–702, 2004.

[18] F. Pulvermüller, M. Huss, F. Kherif, F. M. del Prado Martin, O. Hauk, and Y. Shtyrov, "Motor cortex maps articulatory features of speech sounds," in *Proc. National Academy of Sciences of the United States of America*, vol. 103, 2006, pp. 7865–7870.

[19] G. B. Cogan, T. Thesen, C. Carlson, W. Doyle, O. Devinsky, and B. Pesaran, "Sensory-motor transformations for speech occur bilaterally," *Nature*, vol. 507, pp. 94–98, 2014.

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

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

[22] 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, pp. 2553–2557.

[23] S. Amano and T. Kondo, "Estimation of mental lexicon size with word familiarity database," in *Proc. ICSLP*, 1998, pp. 2199–2122.

[24] M. Garman, *Psycholinguistics in Series of Cambridge Textbooks in Linguistics*. Cambridge University Press, 2008.

[25] 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. Spoken Language Technology*, 2021, pp. 698–704.

[26] A. Favaro, C. Motley, T. Cao, M. Iglesias, A. Butala, E. S. Oh, R. D. Stevens, J. Villalba, N. Dehak, and L. Moro-Velázquez, "A multi-modal array of interpretable features to evaluate language and speech patterns in different neurological disorders," in *Proc. Spoken Language Technology*, 2023, pp. 532–539.

[27] N. Nakanishi, N. Minematsu, and T. Kunihara, "Effects of English shadowing training using unlearned passages on listening and speaking skills," *Language Education & Technology*, vol. 59, pp. 77–105, 2022.

[28] T. Kunihara, C. Zhu, 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, pp. 1303–1307.

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

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

[31] N. Nakanishi, *Introduction of English pronunciation with imported and Japanized words*. Hitsuji Publishing Company, 2022.

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

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

[34] *JACET8000*, https://j-varg.sakura.ne.jp/download/index.html.

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