INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Shadowability Annotation with Fine Granularity on L2 Utterances and Its Improvement with Native Listeners' Script-shadowing

Zhenchao Lin[1], Ryo Takashima[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

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

Abstract

Language teachers often claim that the goal of speech training should be intelligible enough pronunciations, not native-sounding ones, because some types of accented pronunciations are intelligible or comprehensible enough. However, if one aims to provide a technical framework of automatic assessment based on intelligibility or comprehensibility, s/he has to be faced with a big technical challenge. That is collection of L2 utterances with annotations based on these metrics. Further, learners always want to know which parts (words, morphemes, or syllables) in their speech should be corrected. This means that data collection needs a valid method of intelligibility annotation with fine granularity. In our previous studies, a new metric of *shadowability* was introduced, and it was shown experimentally to be highly correlated to perceived intelligibility or comprehensibility as well as it was explained theoretically to be potential to give annotations with fine granularity. In this paper, shadowability annotation with fine granularity is examined experimentally, and a new and more valid method of collecting shadowing utterances is introduced. Finally, we tentatively derive frame-based shadowability annotation for L2 utterances.

Index Terms: Speech assessment, intelligibility, shadowability, annotation, shadowing and script-shadowing, DTW

1. Introduction

Language learners often have foreign accents, many of which are transferred from their L1, and therefore non-native pronunciations can vary due to diversity of learners' L1 [1]. Among these diverse pronunciations, it is true that some types of foreign accents are accepted easily by listeners. Probably due to this, it seems that researchers' attention has been shifted from assessment based on nativeness to that based on intelligibility or comprehensibility [2, 3, 4, 5]. In the latter criterion, more attention should be paid to listeners' behaviors rather than to speakers' (learners') manners of speaking. When one focuses on listeners, s/he will find listener-based sources of diversity, one of which is listeners' L1. Major languages such as English, Chinese, Spanish, etc. are learned by learners of different L1s, and if one wants to discuss how a specific type of pronunciation is accepted by a listener, s/he has to take into good account that listener's profile, such as L1, L2, age, gender, background knowledge, etc. For that, we claim that we should have a valid and easy method for various listeners to assign annotations to L2 utterances in terms of intelligibility or comprehensibility.

Some methods may make it possible to collect L2 utterances with such annotations assigned by various listeners, but learners desire annotations with fine enough granularity. This is because they always want to know which parts (words, morphemes, or syllables) of their utterances should be corrected. In [2], subjective annotation on comprehensibility was asked for listeners to make to each utterance. We consider that granularity is not fine enough practically. In [6], immigrants' English utterances were presented to American listeners, who were asked to repeat the utterances. The repetition voices were transcribed and it was examined how many of intended words were repeated correctly. This can calculate word-based intelligibility but repetition was conducted always after listening. We consider that the obtained intelligibility scores are good in granularity but they are offline scores by observing listeners not while listening but after listening. Online and objective observation of listeners are possible by using physiological sensors [7, 8, 9], where pupillometry and EEG (Electroencephalogram) were used to quantify cognitive load and listening efforts. This approach is good at online observation but the cost is inevitably high and we have to say that it is not practical enough for education. In our previous studies [10, 11, 12, 13], a low-cost, easy, online, and pedagogically-valid method to observe listeners' behaviors of comprehension and to derive scores highly correlated to intelligibility or comprehensibility was proposed. In these studies, a new metric of *shadowability* was examined experimentally.

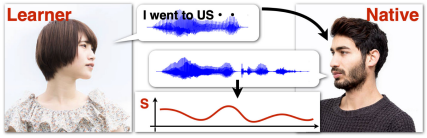

Figure 1: *Reverse form of shadowing*

2. Shadowability-based annotation

In [10, 11], native listeners were asked to shadow given L2 utterances, i.e. repetition of the L2 utterances while listening, but they were required not to mimic learners' accented pronunciations, but just to reproduce in their own native accent what they just heard. Figure 1 shows this reverse form of shadowing, where S means shadowability [14]. Smooth shadowing is possible even by native listeners when and only when listening comprehension is easy and quick [15]. In [10, 11], delay of shadowing and brokenness of articulation were automatically calculated as shadowability. Delays were obtained from alignment between an L2 utterance and its native shadowing, and brokenness of articulation was calculated as a sequence of GOP (Goodness of Pronunciation) scores [16] from the native shadowing, not from the L2 utterance. They were shown to be highly correlated to perceived intelligibility.

http://dx.doi.org/10.21437/Interspeech.2020-2550

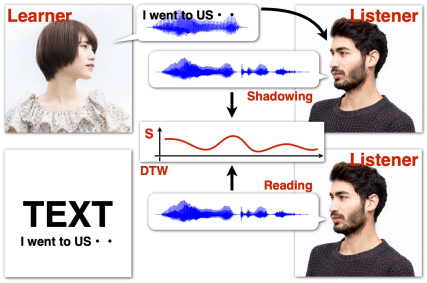

Figure 2: *Comparison between shadowing and reading*

In [13], an improved method of quantitative calculation of shadowability was proposed. After a native listener shadowed an L2 utterance, which is often a read-aloud sentence by a learner, the listener was asked to read the sentence aloud by viewing, shown in Figure 2. Reading is the most prepared speech while shadowing is the least prepared speech. By comparing the two utterances via Dynamic Time Warping (DTW), a sequential data of brokenness of articulation can be obtained on the DTW path between the two. While [13] showed that the DTW-based shadowability is more highly correlated to shadowers' perception than the GOP-based one [10, 11], experimental verifications were made only for utterance-based annotation.

In this study, DTW-based annotation of shadowability is tested again but with finer granularity. Word-based subjective scores are compared to word-based objective scores. Further, we introduce an improved method of data collection, where not reading but script-shadowing is asked for listeners to conduct after they shadow L2 utterances. With this, easy-to-understand presentation of shadowability is made possible. Finally, frame-based and valid annotation on L2 utterances is examined.

3. Word-based shadowability annotation

3.1. Word-based annotation on L2 utterances

In Figure 2, we have three utterances, a reading from a learner, a shadowing from a shadower (native listener), and a reading from the same shadower. In [13], a sequence of local DTW distances were calculated along the DTW path obtained between the two utterances from the shadower. Although this sequential annotation is surely related to the learner's utterance presented to the shadower, we still don't know what kind scores should be assigned to individual words, syllables, phonemes, even frames in the learner's utterance. The sequential annotation obtained in [13] need to be further processed for annotation.

Frame-based assignment will be technically possible, but probably impossible by humans. To what kind of small units, can human raters assign scores reliably? In the experiments, because we compare machine scores and human scores for assessment, and because non-expert native shadowers will have difficulty in assessing even syllable-based units, we decided to conduct experiments using word[1] as basic unit for annotation.

3.2. Data collection [12]

60 Vietnamese learners of Japanese and a professional Japanese narrator were asked to read aloud texts from Japanese textbooks. Among the 60 learners, 27 learners had learned Japanese just for one year and the other 33 learners had learned for two or three years. A part of the recordings, about 800 utterances, were

[1]Strictly speaking, the adopted unit was *bunsetsu*, a word concatenated with a post-positional word of 1-mora or 2-mora length.

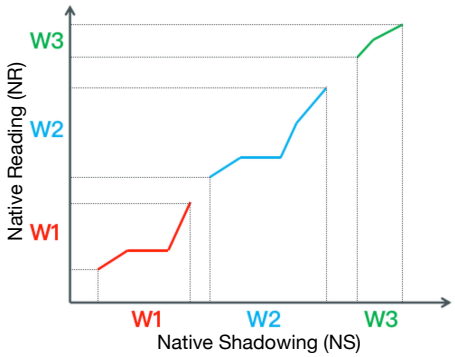

Figure 3: *Word-based DTW between native shadowing (NS) and native reading (NR)*

used in the shadowing experiment. Two Japanese female adult shadowers participated in the experiment, who had not taken any phonetic training or teacher training. They were asked to conduct the following tasks for each of the utterances, a) listen-and-shadow an utterance in their own native accent, and b) read aloud the text of that utterance.

For the current study, the two shadowers were asked again to conduct the following additional task, c) listen to both the shadowing and the reading separately for each word, and rate each word segment in terms of smoothness of shadowing. Here, a four-level scale was used and the four levels indicate (1) totally broken, (2) broken, (3) partially broken, and (4) smoothly shadowed. Word-based score assignment was done only once, but the shadowers were allowed to listen to the recordings repeatedly. These scores are used as word-based subjective scores of shadowability in the following sections.

3.3. Experiments

There are two types of shadowability annotations that can be derived word by word from the DTW alignment between native reading (NR) and native shadowing (NS). One is a sequence of the DTW local distances on the DTW path, which corresponds to brokenness of articulation in shadowing. The other is amount of time required to shadow each word in learner reading (LR), which is calculated by comparing the length of a word in LR and that of its corresponding word in NS. In good and synchronous shadowing, the two lengths are similar but if a word in LR is difficult to understand, then, the length of the corresponding word in NS becomes longer. To detect word boundaries, forced alignment was applied both to LR and NS.

Following [14], posteriorgram was adopted as speech representation and any utterance was represented as a sequence of phoneme-posterior vectors. Here, the CSJ-based KALDI recipe [17, 18] was used to train Japanese DNN acoustic models. Unlike [14], in this study, DTW is always conducted within the same speaker, i.e. between NS and NR. Taking this condition into account, acoustic representation of MFCC was also tested. The local distance between frames was calculated as the Bhattacharyya distance with posteriorgram and as the Euclid distance or the Cosine distance with MFCC.

In [13], DTW was conducted between an overall reading and an overall shadowing from the same native speaker, where pauses were removed in advance to reduce alignment errors. In this experiment, however, since a subjective score was assigned to each word not to an entire utterance, DTW was also conducted separately for each word as in Figure 3, where three paths are drawn for three word segments. For detecting word

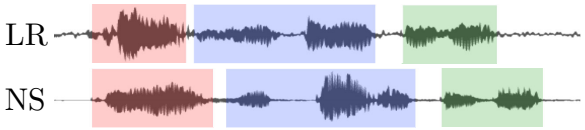

Figure 4: *Word-based temporal lengthening between LR and NS*

boundaries, forced alignment was performed both on NR and NS. The average of the local distances within each word was defined as objective score for that word, which should be assigned to the corresponding word in LR.

Temporal lengthening or shortening in shadowing is represented by the ratio of the length of each word in LR to the length of the corresponding word in NS, shown in Figure 4. For easy comparison, the starting time of LR and that of NS are positioned at the same time index. The first word in NS is longer, while the other two words are similar to those in LR in length.

3.4. Results and discussion

Table 1 shows correlations calculated between the subjective scores and each kind of the objective scores separately for each of the two shadowers, HS1 and HS2. Correlations calculated by dealing with the two shadowers together are also shown. In the table, Posteriorgram shows the highest correlations while MFCC unexpectedly shows low correlations. After the experiments, we found that recording of NSs and NRs were not done in a single session [12]. All the NSs were recorded on a day and all the NRs were recorded on another day. The recording equipment was shared in both recordings, but the recording room was different. This conditional gap may have influenced the experimental results. As for consecutive recording of NS and NR, its technical and pedagogical validity will be discussed later, where effectiveness of MFCC is examined again.

Generally speaking, correlations between subjective scores and objective scores tend to be high when a score is assigned holistically to each learner by using his/her utterances together, i.e. learner-based annotation. As the unit of annotation becomes smaller such as one sentence, one phrase, one word, one syllable, and one phoneme, correlations tend to be smaller [19]. Even in these cases, correlations can become larger when averaged scores are used over multiple raters. This is because inevitable deviations in subjective assessment can be reduced. In the experiment here, we took what is supposed to be the minimum unit for non-expert native speakers, i.e. word. Further, as we pointed out in Section 1, when listener-based diversity is taken into account, averaging operations over listeners may not be adequate. This is why we did not calculate the averaged scores over the shadowers but calculated correlations by using their scores as independent data. In spite of these difficult conditions, Posteriorgram shows very high correlations in Table 1.

It should be much noted that Posteriorgram's correlations in Table 1 are higher than those obtained in our previous study [13], where posteriorgram-based DTW was conducted between NS and NR for utterance-based annotation. The reason of higher correlations for shorter units is considered to be due to uniqueness of native shadowing. In shadowing an L2 utterance, it is likely that many words are shadowed smoothly while a few others are not. In this case, brokenness of articulation is suddenly raised. Distinction between these words is impossible for utterance-level annotation but easy for word-based annotation.

In Table 1, the averaged score of phoneme-based GOPs (pGOPs) calculated for each NS word and that for each LR word are also used to calculate their correlations to word-based subjective scores. Their correlations are by far lower than Posteriograms' correlations. In [10], the averaged score of pGOPs was calculated for each NS utterance and it showed a high correlation (0.73) to subjective scores. In [10], an objective score was obtained as the averaged score among 27 shadowers' shadowings and a subjective score was also obtained as the averaged score among 27 shadowers' judgments. In [12], the same data was re-analyzed without averaging operations, and the resulting correlation was found to be low (0.50). The averaging operation is powerful to increase correlations, but Posteriorgram's correlations in Table 1 were obtained without averaging operations over raters. In the next section, we introduce a small change in the data collection protocol of shadowing and reading.

Table 1: *Correlations of word-based subjective scores and objective scores for each of the two shadowers*

Feature	HS1	HS2	both
MFCC_E	0.454	0.446	0.457
MFCC_C	0.169	0.128	0.144
Lengthening	0.430	0.329	0.375
Posteriorgram	**0.709**	**0.797**	**0.734**
Posteriorgram [13]	0.600	0.620	—
pGOP (NS)	0.579	0.528	0.572
pGOP (LR)	0.108	0.205	0.123

4. Pedagogically-valid data collection through shadowing and script-shadowing

4.1. Three inevitable problems and a simple solution

The revserse-shadow-and-read method has been proven to be effective with finer granularity. Here, a shadowing is treated as the least prepared speech and a reading is as the most prepared speech. If both utterances are similar, easiness and quickness of understanding should be high. It is a simple principle.

In this method, however, we found three inevitable problems. 1) Different shadowers adopt different shadowing strategies. Some shadowers try to minimize delay of shadowing, paying less attention to articulation, and others try to maximize articulation, paying less attention to delay. When both types of shadowers read in a similar reading style, even in the case that DTW-distances between NS and NR are different between the two types of shadowers, they sometimes assign similar subjective scores. 2) Reading styles can vary among shadowers. L2 utterances are often slow, and when native shadowers read the text, some of them read it quickly. Quick phonation often results in inarticulate phonation even when natives read. This causes a bias when calculating shadowability. 3) Recording NS and NR is done independently. When the two utterances are presented as waveforms to teachers and learners, it is difficult to interpret acoustic gaps between the two utterances.

In this section, we introduce a small change of the data collection protocol into the method examined in the previous section. Although our solution is not a technical solution, its effectiveness is very high. In the previous section, a shadowing and a reading were viewed as the least prepared speech and the most prepared speech, respectively. To solve the above problems simultenously, we realized that comparison should be made not between a shadowing and a reading, but between a shadowing and the best shadowing. The best shadowing can be obtained by asking a shadower to shadow repeatedly or asking a shadower to shadow with transcript given. Here, we took the second option and, in applied linguistics, this form of shadowing is often called script-shadowing, shown in Figure 5.

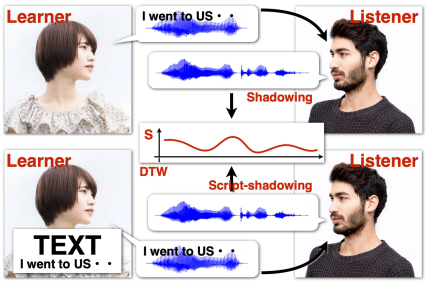

Figure 5: *Comparison bet. shadowing and script-shadowing*

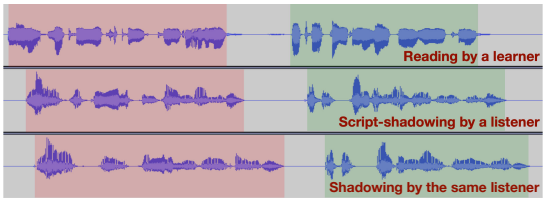

Figure 6: *An example triplet of LR, NSS, and NS*

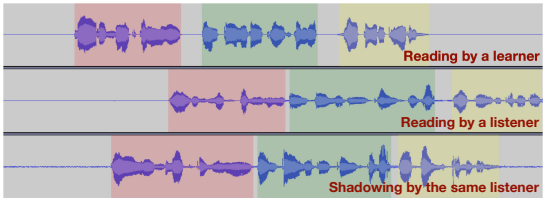

Figure 7: *An example triplet of LR, NR, and NS*

To analyze utterances of shadowing and script-shadowing, we prepared two datasets. One contains the two types of shadowing utterances from native listeners of English, to whom 30 Japanese English utterances were presented. The other contains the two types of utterances from native listeners of Japanese, to whom 30 Chinese Japanese utterances were presented.

4.2. Easy-to-understand presentation of utterances

Figure 6 shows a typical example of a reading from a learner (LR), a script-shadowing from a native shadower (NSS), and a shadowing from the same shadower (NS). Since all the three utterances share the same time axis, the temporal structure of these utterances can be directly compared. For example, as NSS can be viewed as the best shadowing, delay in NSS is short but in NS, it becomes longer. Phrase boundaries are manually visualized with different colors. Without those illustrations, however, even learners can understand that the first phrase become longer in NS compared to that in LR and NSS. This implies that some parts in the first phrase in LR are difficult to understand quickly, and by listening to NS, learners can get to know which parts reduced comprehensibility. Figure 7 shows an example of LR, NR, and NS. LR and NS share the time axis but NR was recorded independently of LR and NS. Then, NR cannot be compared to LR and NS visually and directly. Pedagogically speaking, Figure 6 is by far more informative than Figure 7.

4.3. Speaking rate control realized in script-shadowing

In Figure 6, the length of each phrase of NSS tends to be similar to that in LR because NSS is basically synchronous reading with LR. Before collecting NS and NSS for Japanese English utterances, NR was also recorded independently. Table 2 shows the ratio of utterance-based NR/LR and that of NSS/LR for each of the native shadowers. For every shadower, NSS/LR becomes closer to 1.0. By asking native shadowers to script-shadow, we can obtain utterances temporally aligned to LR.

Table 2: *Ratios of phrase lengths in NR, NSS, and LR*

	S1	S2	S3	S4	S5	S6	S7
NR/LR	1.40	0.76	0.75	0.82	0.94	1.18	0.80
NSS/LR	1.20	0.81	1.01	0.98	0.99	1.02	0.99

Table 3: *Correlations of posterior-based phonemic distances and purely acoustic distances*

HS1	HS2	both	CR
0.661	0.627	0.658	0.822

CR = Consecutive Recording of NS and NSS.

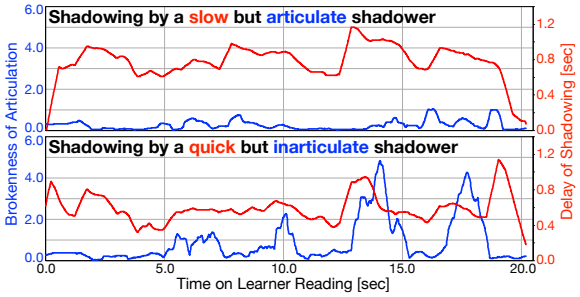

Figure 8: *Frame-based annotation of shadowability*

4.4. Acoustic comparison between consecutive recordings

In Section 3.2, NR and NS were recorded in different rooms. and in this section, NSS and NS were recorded consecutively in the same room. Posterior-based DTW was conducted for pairs of NR and NS and for pairs of NSS and NS. At each node on the obtained DTW paths, the MFCC distance between the two corresponding frames was calculated. Here, the MFCC distance was calculated as weighted Euclidean distance. Table 3 shows word-based correlations between posterior-based distances and MFCC distances. By consecutive recording, posterior-based DTW can be replaced by MFCC-based DTW. The former requires DNN models and therefore this approach is difficult to be applied to minority languages. Consecutive recording guarantees the effectiveness of our approach to any language.

4.5. Toward frame-based shadowability annotation

Frame-based annotation of shadowability is calculated and visualized tentatively. Here, moving average is conducted on every 0.5 sec segment with 10 msec shift for the Posterior-based DTW scores. Frame-based delays between LR and NS are also averaged in a similar way. Figure 8 shows shadowability graphs for a single LR, shadowed by two shadowers. One shadower shadows smoothly but with long delays, while the other shadows not so smoothly but with small delays. In the figure, a strategic difference of shadowing is observed between shadowers.

5. Conclusions

Our proposed method of NS and NR for annotating L2 utterances was examined experimentally with finer granularity. Results indicated high effectiveness of our proposal even when it is applied to word-based annotation without averaging operations over raters. Further, an improvement was realized by replacing NR by NSS. With NSS, the collected utterances can be more informative to learners. Finally, frame-based annotation of shadowability was presented although this is still tentative.

6. References

[1] D. Birdsong, "Nativelikeness and non-nativelikeness in L2A research," *International Review of Applied Linguistics in Language Teaching*, vol. 43, no. 4, pp. 319–328, 2005.

[2] M. J. Munro and T. M. Derwing, "Foreign accent, comprehensibility, and intelligibility in the speech of second language learners," *Language Learning*, vol. 45, no. 1, pp. 73–97, 1995.

[3] ——, "The functional load principle in esl pronunciation instruction: An exploratory study," *System*, vol. 34, pp. 520–531, 2006.

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

[5] T. M. Derwing, "Comprehensibility," in *The TESOL Encyclopedia of English Language Teaching*, J. I. Liontas, Ed. John Wiley & Sons, Inc, 2018, pp. 570–584.

[6] J. Bernstein, "Objective measurement of intelligibility," in *Proc. ICPhS*, 2003, pp. 1581–1584.

[7] 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.

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

[9] 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.

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

[11] 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. Spoken Language Technology*, 2018, pp. 971–978.

[12] S. Ando, Z. Lin, T. Trisitichoke, Y. Inoue, F. Yoshizawa, D. Saito, and N. Minematsu, "A large collection of sentences read aloud by vietnamese learners of japanese and native speaker's reverse shadowings," in *Proc. O-COCOSDA*, 2019, pp. 1–6.

[13] Z. Lin, Y. Inoue, T. Trisitichoke, S. Ando, D. Saito, and N. Minematsu, "Native listeners' shadowing of non-native utterances as spoken annotation representing comprehensibility of the utterances," in *Proc. SLaTE*, 2019, pp. 43–47.

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

[15] 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.

[16] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, no. 1, pp. 95–108, 2001.

[17] 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.

[18] "Corpus of Spontaneous Japanese(CSJ)." [Online]. Available: https://pj.ninjal.ac.jp/corpus_center/csj/en/

[19] S. Nakagawa, K. Mori, and N. Nakamura, "A statistical method of evaluating pronunciation proficiency for english words spoken by japanese," in *Proc. INTERSPEECH*, 2003, pp. 3193–3196.