SLaTE 2019: 8th ISCA Workshop on Speech and Language Technology in Education
20-21 September 2019, Graz, Austria

Native Listeners' Shadowing of Non-native Utterances as Spoken Annotation Representing Comprehensibility of the Utterances

Zhenchao Lin, Yusuke Inoue, Tasavat Trisitichoke, Shintaro Ando, Daisuke Saito, Nobuaki Minematsu

Graduate School of Engineering, The University of Tokyo
{zhenchaolin,inoue0124,tasavat,s_ando,dsk_saito,mine}@gavo.t.u-tokyo.ac.jp

Abstract

Recently, researchers' attention has been paid to pronunciation assessment not based on comparison between learners' utterances and native models, but based on comprehensibility of the utterances [1, 2, 3]. In our previous studies [4, 5], native listeners' shadowing was investigated and shown to be effective to predict comprehensibility perceived by listeners (shadowers). In this paper, native listeners' shadowings are viewed as spoken annotations that can represent comprehensibility. In [4, 5], to predict comprehensibility of a non-native utterance, the GOP scores of its corresponding native listeners' shadowings were calculated by using a DNN-based ASR front-end. Generally speaking, annotations are prepared manually and, even when some techniques are used for annotations, only stable and reliable techniques should be used. In this paper, a simpler, stabler, and more reliable method to derive comprehensibility annotations was proposed. After native listeners' shadowing, they are asked to read aloud the sentence intended by the learner. Reading is the most prepared speech and shadowing is probably the least prepared speech. DTW between the two utterances is supposed to be able to quantify and predict comprehensibility or shadowability perceived by the shadowers. In experiments, DTW between shadowings and readings shows higher correlation than the GOP scores of shadowings.

Index Terms: Pronunciation assessment, comprehensibility, shadowing and reading, GOP, DTW

1. Introduction

In previous studies aiming at automatic pronunciation assessment [6, 7, 8], comparison between learners' utterances and a native model of pronunciation was often made. However, some types of foreign accents hardly reduce smoothness of communication [1, 2, 3], and many teachers claim that the goal of pronunciation training is an intelligible-enough or comprehensible-enough pronunciation, not a native-sounding one [1]. Automatic pronunciation assessment should be realized not based on native-likeness but based on intelligibility or comprehensibility.

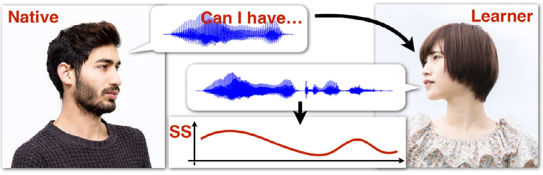
In applied linguistics, intelligibility of an utterance indicates how many linguistic units such as words can be identified correctly. Degree of intelligibility of an utterance can be measured objectively by asking native listeners to transcribe or repeat that utterance after listening to it [9, 2]. Comprehensibility of an utterance means how smoothly listeners can understand the content of that utterance, and degree of comprehensibility has been often quantified by listeners' subjective rating [2]. Listening effort [10] and cognitive load [11] seem to be strongly related to comprehensibility, i.e. smoothness of understanding.

Intelligibility was measured objectively in [9, 12], where English spoken by immigrants to the USA [9] and English spoken by Japanese college students [12] were presented to native listeners on the telephone. They were asked, after listening, to repeat what they heard. Their repetitions were transcribed manually by technical staff and these transcriptions were used as intelligibility annotations representing which word segment in the non-native utterances is perceived by native listeners with how much accuracy [13, 14]. However, the authors consider that this approach lacks in its scalability. Collection of transcriptions requires cost and time and therefore, this approach can be said not to be practical enough to increase the amount of annotations.

Lack of annotation is a well-known problem for CALL (Computer-Aided Language Learning) studies [15]. It seems also to be a general problem for machine learning studies and DNN-based artificial intelligence studies [16]. Although a large number of non-native speech databases are available [17], it is not rare that researchers cannot find annotations or labels in the databases which the researchers need for their specific purposes. One reason for this is annotating non-native utterances often requires expert phoneticians or teachers especially when small units such as syllables and phonemes are selected as the unit of annotation. In the case that holistic annotation is needed, such as rating easiness of understanding of a given utterance, annotation can be done by ordinary and non-expert native listeners. In this case, however, the resolution of annotation is generally low and a score is often given to an entire utterance or a set of utterances, not to individual words or syllables.

How to collect a huge amount of annotation based on intelligibility or comprehensibility of non-native utterances, even with high enough resolution and by not asking expert phoneticians and teachers? In other words, how to make it feasible to collect those annotations on a cloud sourcing infrastructure? One possible solution is collecting data related to listeners' behaviors or responses while listening to non-native speech. In [18, 19, 10], EEG (electroencephalogram) recordings were made from listeners and listening efforts were discussed quantitatively and in [11], eye-trackers were used to measure the size of pupils to predict the magnitude of cognitive load when listening. These features are strongly related to comprehensibility and obtained recordings may give us a sequential data of comprehensibility, but physiological sensing requires expensive devices. This is why the authors consider that physiological sensing is impractical and very difficult to be used on a cloud sourcing infrastructure. Our previous studies [4, 5] showed another possibility only with an inexpensive device, that is a microphone, and listeners' responses were observed acoustically. Native listeners were asked to shadow given non-native speech and smoothness of their shadowing was calculated acoustically. Since smooth shadowing always requires smooth understanding [20], comprehensibility scores were more highly correlated to GOP scores of native listeners' reverse shadowings than those of non-native utterances that were presented to shadowers. In this paper, those shadowings are viewed as spoken annotations on comprehensibility and it is discussed how to obtain those annotations more adequately, or how to calculate smoothness of

10.21437/SLaTE.2019-8



SS means smoothness of shadowing.
Figure 1: *Conventional form of shadowing*

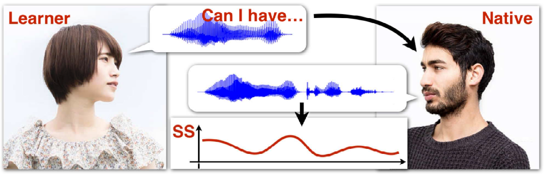


Figure 2: *Reverse form of shadowing*

shadowing technically in a more stable and reliable way.

2. Related works

2.1. Conventional form of shadowing

Shadowing is a special type of listen-and-repeat practice, where a listener has to repeat a given utterance as simultaneously as possible, shown in Figure 1. Shadowing was originally introduced as a practicing strategy for simultaneous interpreters since it includes not only speaking and listening but also understanding a given speech. Recently, researches and teachers have shown that shadowing is also effective for second language learning [21, 22, 23]. Conversation is generally a speech activity where three processes of speaking, listening, and understanding are overlapped. Practically speaking, conversation is a multi-task speech activity, and shadowing is used in classrooms as it can put learners effectively in this multi-task situation. In [8], some of the authors proposed a DNN-based technique to predict smoothness of shadowing, SS for short in Figure 1.

2.2. Reverse form of shadowing

In [4, 5], a novel method of predicting comprehensibility of an utterance was proposed, that is native listeners' reverse shadowing and it does not require any special device like EEG or eye-tracker. In the conventional form of shadowing, native utterances are presented to learners, who are shadowers. In reverse shadowing, learners' utterances are presented to native shadowers, shown in Figure 2. Here, shadowers are asked not to imitate accented pronunciations but to reproduce what was said in their own native pronunciation. Since smooth shadowing always requires smooth understanding [20], smoothness or brokenness of natives' shadowing, which was acoustically measured, was examined and shown to be effective to predict comprehensibility subjectively rated by shadowers.

2.3. Comprehensibility prediction from natives' shadowing

Two speech features used were used in [4, 5] to predict comprehensibility. They are accuracy of articulation and delay of shadowing. As for the former, we used Goodness Of Pronunciation (GOP) measure [7, 6, 8]. GOP is a widely-used baseline speech feature in pronunciation assessment studies and, when GOP is applied to an L2 utterance, it represents how similar that utterance is to the model pronunciation in terms of articulation (phoneme generation). GOP is theoretically defined as phoneme-based posterior $P(c_i|o_t)$, where o_t is a speech feature observed at time t, and c_i is phonemic class i. In Figure 2, after forced alignment performed on the native shadowing with the string of phonemes intended by the learner, $P(p_t|o_t)$ is averaged over the entire duration of a given phonemic segment, where p_t is the phoneme shadowed at time t. The GOP of the k-th segment is calculated as

$$\text{GOP}(k) = \frac{1}{D_k}\sum_{t\in x} P(p_t|o_t), \qquad (1)$$

where D_k is the frame-based total duration of the k-th segment. For a shadowing utterance with K segments, the averaged GOP score over $\{\text{GOP}(k)\}$ is calculated as shadowability score for that utterance.

In [4, 5], Japanese sentences read by Vietnamese learners and native listeners' shadowings were used for analysis, and $P(p_t|o_t)$ was calculated by a DNN-based front end of a Japanese speech recognizer, trained with CSJ [24]-based KALDI [25]. The GOP scores from natives' shadowings were shown to be very highly correlated with comprehensibilities subjectively judged by the native shadowers.

As for delay of shadowing, in [4, 5], by comparing a forced alignment result of a Vietnamese-Japanese (VJ) utterance and that of its native reverse shadowing (RS), the temporal gap between every pair of phoneme boundaries was obtained between the two utterances. The phoneme-based temporal gaps obtained from the two utterances were averaged to define delay of shadowing between the two. Generally speaking, shadowing is performed with a delay of about 1 second to a presented utterance.

3. Proposed method

3.1. Problems in our previous studies

The authors consider that it is very natural to view native listeners' shadowing as a kind of annotation assigned to a given non-native utterance, which can characterize comprehensibility of that utterance even in the form of sequential data, not a single score. As was shown in [4, 5], non-expert and ordinary native listeners can be adopted as shadowers, but the authors can point out two drawbacks when we simply use DNN-based GOP scores to represent comprehensibility dynamics.

Generally speaking, annotations or labels are given manually to speech samples. When some techniques are used for annotation, they should be stable and reliable techniques. DNN-based ASR can work more stably and reliably than HMM-based ASR, but DNN-based ASR still needs adaptation techniques with respect to speaker identity, speaking style, recording environment, etc. This fact implies that DNN-based GOP scores can be unstable and unreliable in some specific situations.

The other drawback is more crucial. In [4, 5], the target language of learning was Japanese, which is not an international language, and in this case, shadowers should be native speakers of Japanese. If we adopt English as target language, as it is used internationally, some learners want to know how comprehensible their utterances are to non-native speakers of English. In this case, we can ask non-native listeners to shadow. Even when their shadowing is very smooth, however, their utterances are often accented, which easily reduces GOP scores if the DNN-based acoustic models are trained with a native speech corpus. If non-native acoustic models are available separately for each of native languages, the above problem may be able to be solved, but this is very impractical. Further, even native shadowers may be rejected as shadower if their pronunciations are regionally accented. To use listeners' shadowing as spoken annotation effectively, a different method of calculating smoothness of shadowing is needed, which should be stabler.

3.2. Proposed method

In this paper, we solve the above problem just by introducing another simple speech task to shadowers. When listeners shadow a given utterance, they do not refer visually to the sentence that the learner read aloud. Then, after they shadow, we present the sentence visually and ask them to read it aloud. The authors consider that reading is the most prepared speech and shadowing is probably the least prepared speech, or hastened speech. If smooth or quick understanding is possible enough while shadowing, the shadowing speech will become acoustically closer to the reading speech. Dynamic Time Warping (DTW) between the two types of speech gives us the optimal path, on which a sequence of local distances can be viewed as a sequence of comprehensibility. Further in this method, DTW is conducted within the same speaker, with the same microphone, and in the same room. The authors can claim that this is the best recording condition for utterance comparison using DTW.

4. Data collection [26]

4.1. Collection of Vietnamese Japanese readings

From 60 Vietnamese learners of Japanese, we collected two kinds of reading utterances, readings of sentences in Japanese textbooks and those of the learners' own essays. The period of learning Japanese was one year for 27 learners and two to three years for the other 33 learners. For textbook reading, recording was done by a unit of phrase and every recording was so long as a phrase. For essay reading, recording unit was not fixed and many learners recorded sentence by sentence, but some others recorded their whole essay in a single recording session. A part of the recordings was used for collecting native listeners' shadowing utterances, which is explained below.

4.2. Collection of natives' shadowings and readings

Two native speakers participated in our experiments and each of them shadowed a different set of 800 utterances of Vietnamese-Japanese (VJ). After shadowing each VJ utterance, the shadowers rated subjectively degree of comprehensibility. After the entire shadowing experiment, we recorded the shadowers' reading utterances. For each shadower, 400 sentences that the Vietnamese learners used for recording, i.e. a half amount of data used for shadowing recording, were visually presented to the shadower, who was asked to read aloud the sentences. After reading aloud each sentence, both of the reading utterance and its corresponding shadowing utterance were presented to the shadower through headphones so that s/he could rate degree of shadowability in a seven-degree scale, i.e. how correctly or incorrectly articulation was performed in shadowing. [1]. In the following section, we compare correlations of the GOP scores of native listeners' shadowings to the shadowability scores and those of the DTW scores between shadowing and reading to the shadowability scores.

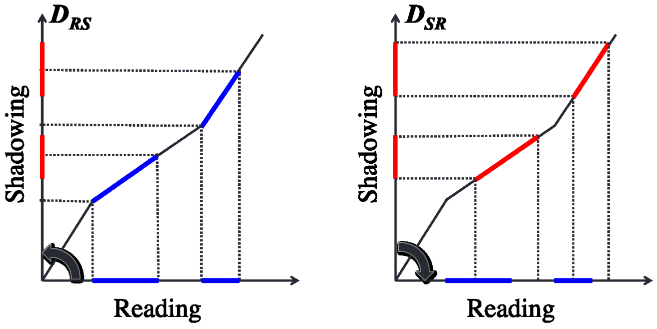


Figure 3: *Two kinds of DTW scores, D_{RS} and D_{SR}*

5. Experiments

5.1. Detailed procedures of DTW

For DTW between shadowings and readings, all the utterances were converted to their posteriograms with a DNN-based ASR front end [4]. The most problematic thing in the DTW is that readings and shadowings often have pauses at different positions within the utterances. In readings, pauses are intentionally inserted at punctuations or phrase boundaries in sentences, so that the utterances will become more natural and clear. Nevertheless, in shadowings, pauses are often found not at syntactic boundaries but at positions where listeners' understanding process did not work smoothly and they had to wait to continue shadowing. This irregular pausing is mainly due to low shadowability of original learners' utterances. To calculate shadowability scores objectively from the DTW path, pauses in mismatched positions between readings and shadowings have to be handled in a proper way.

Further, while some words are missing in shadowing utterances, words that are not in readings can sometimes be found in shadowings, e.g. repetitions and unintentionally produced words of surprise such as what or hmm. To handle these phenomena adequately, the following procedure was examined.

Comparison between a shadowing and its reading was done via. DTW and accumulated distances are calculated only on speech segments. We prepared two types of distances, D_{RS} and D_{SR}, shown in Figure 3. In D_{RS}, reading was used as reference and speech segments were detected from the reading, that are drawn in blue in Figure 3. The DTW paths for those speech segments were used to calculate the accumulated distance. In D_{SR}, shadowing was used as reference and speech segments were detected from the shadowing, that are drawn in red in the figure. The DTW path for those speech segments were used. In either case, the accumulated distance was normalized by the number of speech frames in reading or shadowing.

Another question is how to detect pauses in a specific utterance. In this paper, two methods for pause detection are applied. One is based on the result of forced alignment and this method can be applied only when text of learners' speech is available. The other is based on posteriogram and this method can be applied even without text of learners' speech.

In Figure 4, an example of DTW of utterances of speaker HS001 is illustrated. Background color is painted to represent local distances, where darker red means larger distance while deeper blue means smaller. Speech segments found in the optimal path are painted by scattered white dots. The upper figure in Figure 4 is drawn based on D_{RS} and the lower is drawn with D_{SR}. In the upper, it is found that some speech frames in reading are missing in shadowing.

[1] The authors consider that comprehensibility and shadowability are practically similar metrics, but theoretically and strictly speaking, both have a gap between them. In this study, we're interested more in shadowability because we aim at predicting smoothness of understanding via. observing listeners' shadowing behaviors.

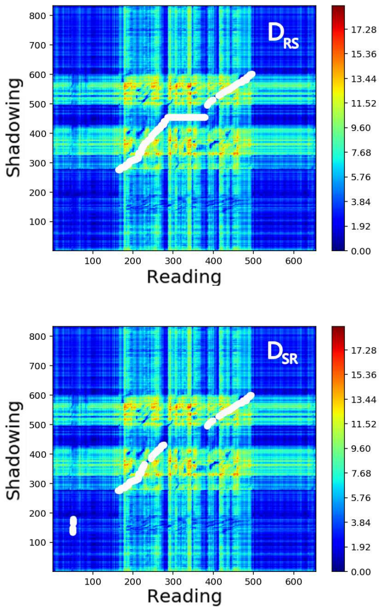


Figure 4: *Two kinds of DTW path, D_{RS} and D_{SR}*

5.2. Results and discussion

For a pair of a shadowing and its reading, two DTW scores are linearly combined as $\alpha D_{RS} + (1-\alpha)D_{SR}$, where α varied from 0.0 to 1.0 with a step of 0.1. When α=1, the combined score is the same as D_{RS} and when it is 0, the score is the same as D_{SR}. Correlations of the combined scores to the shadowability scores are shown in Table 1 for each of the two native listeners, where the upper table shows the correlations with text and forced alignment and the lower shows those without text but with posteriogram. In each table, correlations of GOP scores calculated only on shadowings are also shown.

The tables show that D_{RS} tends to have higher correlations than D_{SR}, but in the case of shadower HS002 with text, D_{SR} shows a higher correlation than D_{RS}. From this table, we can say that 0.6 will be the best weight for α. It is clearly shown that the scores based on DTW between shadowings and readings have higher correlations than the GOP score calculated only from shadowings but with DNN-based acoustic models. As discussed in Section 3.1, DTW-based scoring has much higher availability as it can be applied to non-native shadowers.

6. Conclusions

In this paper, we discussed some existing problems of the method of native listeners' reverse shadowing and proposed a simple solution so that native listeners' reverse shadowing can be used effectively as spoken annotations. In the proposed method, in addition to shadowings, readings have to be collected from shadowers. However, reading-shadowing DTW experimentally showed to provide scores that are more highly correlated to perceived shadowability than GOP scores calculated from shadowings. Further, availability of the DTW-based comparison is much higher because it can be directly applied to non-native listeners' shadowings and readings. As future work, other features like MFCC (Mel Frequency Cepstrum Coefficients) will be examined for calculating DTW path. It is also interesting to compare shadowing and reading in multiple resolutions, where phoneme-level, syllable-level, and word-level comparisons are done to give hierarchical assessment scores to shadowing-based annotations. With a huge enough number of annotations available, we will attempt to build a virtual shadower, where a learner's utterance will be converted to its shadowability sequence or shadowability scores for each word in the utterance.

Table 1: *Correlation of DTW and GOP to shadowability*

1) with text and forced alignment

α	HS001	HS002
0.0(D_{SR})	-0.52	-0.61
0.1	-0.54	-0.61
0.2	-0.55	-0.62
0.3	-0.57	-0.61
0.4	-0.58	-0.61
0.5	-0.59	-0.60
0.6	-0.60	-0.59
0.7	-0.60	-0.58
0.8	-0.61	-0.57
0.9	-0.61	-0.56
1.0(D_{RS})	-0.61	-0.54
GOP	0.41	0.50

2) without text but with posteriogram

α	HS001	HS002
0.0(D_{SR})	-0.55	-0.60
0.1	-0.56	-0.62
0.2	-0.57	-0.63
0.3	-0.57	-0.63
0.4	-0.58	-0.64
0.5	-0.58	-0.64
0.6	-0.59	-0.64
0.7	-0.59	-0.64
0.8	-0.59	-0.63
0.9	-0.60	-0.63
1.0(D_{RS})	-0.60	-0.62
GOP	0.45	0.55

7. References

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

[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] 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.

[5] 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.

[6] 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.

[7] W. Hu, Y. Qian, F. K. Soong, and Y. Wang, "Improved mispronunciation detection with deep neural network trained acoustic

models and transfer learning based logistic regression classifiers," *Speech Communication*, vol. 67, pp. 154–166, 2015.

[8] 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.

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

[10] 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.

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

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

[13] T. Pongkittiphan, N. Minematsu, T. Makino, and K. Hirose, "Automatic detection of the words that will become unintelligible through japanese accented pronunciation of english," in *Proc. SLaTE*, 2013, pp. 109–111.

[14] T. Pongkittiphan, N. Minematsu, T. Makino, D. Saito, and K. Hirose, "Automatic prediction of intelligibility of english words spoken with japanese accents – comparative study of features and models used for prediction –," in *Proc. SLaTE*, 2015, pp. 19–22.

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

[16] A. Rosenberg, "Speech, prosody, and machines: Nine challenges for prosody research," in *Proc. Speech Prosody*, 2018, pp. 784–793.

[17] "Non-native speech database." [Online]. Available: https://en.wikipedia.org/wiki/Non-native_speech_database

[18] 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.

[19] A. Hahne, "What's different in second-language processing? evidence from event-related brain potential," *Journal of Psycholinguistic Research*, vol. 30, no. 3, pp. 251–266, 2001.

[20] 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 (to appear)*, 2019.

[21] 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.

[22] ——, "Shadowing: Who benefits and how? uncovering a booming efl teaching technique for listening comprehension," *Language Teaching Research*, vol. 20, no. 1, pp. 35–52, 2016.

[23] K. T. Hsieh, D. H. Dong, and L. Y. Wang, "A preliminary study of applying shadowing technique to english intonation instruction," *Taiwan Journal of Linguistics*, vol. 11, no. 2, pp. 43–65, 2013.

[24] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of japanese," in *Proc. LREC*, 2000, pp. 947–952.

[25] 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.

[26] S. Ando, T. Trisitichoke, Y. Inoue, F. Yoshizawa, D. Saito, and N. Minematsu, "A large collection of japanese sentences read aloud by vietnamese learners and native speakers' responsive shadowings," in *Proc. Spring Meeting of Acoustic Society of Japan*, 2019, pp. 111–114.