（『神戸学院大学グローバル・コミュニケーション学会紀要』第 7 号 2022. 3. 31）

# Shadowability of English Learners' Utterances: Comparison with Transcripts Generated by Automatic Speech Recognition

Noriko NAKANISHI [1]
Nobuaki MINEMATSU [2]
Chuanbo ZHU [2]

**Key words**: L2 pronunciation, reverse shadowing, automatic speech recognition

## 1. Introduction

It is desirable for English learners to acquire intelligible pronunciation that enables them to communicate in international oral communication settings (Jenkins, 1998; 2005; Levis, 2005; 2020; Munro, 2008). However, research shows various factors that make English teachers hesitate to conduct pronunciation training in classrooms. A crucial problem is that teachers often have difficulties in identifying learners' pronunciation errors on the spot (Nakanishi, Tam, & Ebihara, 2020). The use of automatic speech recognition (ASR) can be one of the solutions that offer individual, intensive, and anxiety-free pronunciation practice in such classroom situations. However, some disadvantages of ASR-driven pronunciation learning, such as its low recognition rate, have been reported (Coniam, 1999; Derwing, Munro, & Carbonaro, 2000).

This study examined a possibility of implementing reverse shadowing (RS) as another means of English pronunciation practice. RS is an activity in which the L2 learners' recorded utterances are repeated by persons who conduct shadowing (hereafter, shadowers). That is, the roles of the speaker and shadower are reversed from the conventional shadowing (see Kadota (2019) for the method and effects of conventional shadowing). In conventional shadowing practices, L1 utterances are usually presented as a model to L2 learners who try to imitate the model. Conversely, in the case of RS, L2 learners' utterances are presented to international users of English, who play a role of shadower. Here, shadowers are not asked to imitate the learners' accented pronunciations, but to instantaneously reproduce the utterances as perceived. By listening to the reverse-shadowed utterances, the learners can realize how intelligible their utterances were to the shadowers, and how their utterances would likely be perceived in authentic oral communication. In Zhu, Hakoda, Saito, Minematsu, Nakanishi, and Nishimura (2021), shadowers' utterances were used for automatically generating acoustic shadowability scores, which were compared with the intelligibility scores

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

calculated from manual transcription of the same utterances. The results showed high correlations between these two scores.

In this paper, following brief summaries of previous ASR development and studies related to RS, the RS performances of L2 speech will be compared 1) within the shadowers and 2) with the ASR transcriptions. Based on these results, some implications for future research, with the application of RS to educational use in mind, will be suggested.

### 1.1 ASR and L2 Speech

ASR systems were originally developed for practical use such as operating machines, making minutes, and analyzing phone calls. The systems convert input speech into text, referring to acoustic and language models. The acoustic models are for individual phonemes, which are used to represent the pronunciation of each word of the entire lexicon. The language models contain probabilities of words and their transitions (Cucchiarini & Strik, 2018). It should be noted that these two models are generally constructed with speech corpora and text corpora built from a collection of native speakers' data. In other words, performances and behaviors of ASR are not equivalent to those of international speakers of English around the world, who are not necessarily native English speakers.

Attempts to employ ASR for detecting L2 speech pronunciation errors began in the 1990s. Cucchiarini and Strik (2018) introduced some of the studies that had measured the effectiveness of pronunciation training software developed in the early stages. The results lead to critical views on the ASR-based pronunciation training, since those systems were "driven by the possibilities offered by the technology than by pedagogical requirements (p. 558)". Initial claims were that the poor recognition results could reduce the learners' motivation than offer them support, and that the results should be accompanied by appropriate guidance to improve learners' pronunciation (Coniam, 1999; Derwing, Munro, & Carbonaro, 2000; Hincks, 2002; McCrocklin, 2016). Nevertheless, following the improvement of the ASR technology, positive effects of the ASR-driven dictation activities were also reported (McCrocklin, 2019; Mroz, 2018; Neri, Cucchiarini, & Strik, 2003; also see Shadiev, Hwang, Chen, and Huang (2014) for a review of related studies).

In previous studies, the accuracy of ASR performance was often derived from comparisons with human perception of L2 speech. However, the methods of measurement should be taken into consideration. One side of the comparison is how the speech is dictated by the ASR system. As described above, the transcription of ASR depends on the acoustic and the language models, which are usually trained with data derived from native speakers' utterances. The system may not show a good performance with particular L2 utterances, which can still be intelligible in real international communication. The other side of the comparison is how humans perceive L2 speech, quantified by

listeners' subjective or objective judgement. The measurements such as intelligibility, comprehensibility, and accentedness (Munro & Derwing, 1995; Derwing & Munro, 1997) have been widely discussed and cited. The degrees to which the L2 speech is assessed by humans vary depending on the scales (e.g. impressionistic evaluation, word-by-word transcription, overall comprehension). Moreover, in international communication settings, the listeners are not necessarily fluent in the language used. Thus, when comparing the ASR-based transcriptions with human judgment of L2 utterances, it is essential to consider how robust the ASR system is to what kind of contextual factors such as accents in varieties of English, as well as which aspect of human judgement (e.g. intelligibility, comprehensibility, accentedness) is focused.

### 1.2 Reverse Shadowing

The present research started with a question of "how to observe listeners' process of identifying the individual words in a given L2 speech." Observation should be made while listening, not after listening, and it should also be objective and analytic, which is different from subjective and holistic metric such as comprehensibility. One possible way of such observation is monitoring listeners' behaviors based on physiological sensing such as measuring the size of listeners' pupils (Govender, & King, 2018) and/or the Electroencephalogram (EEG) waveforms of their brains (Goslin, Duffy, & Floccia, 2012; Romero-Rivas, Martin, & Costa, 2016). With special equipment for pupillometry and EEG, researchers can obtain the results of observing listeners' process of word identification as temporal and sequential data, which makes analytical discussions possible, but the cost is extremely high.

In Inoue, Kabashima, Saito, Minematsu, Kanamura, and Yamauchi (2018), RS was applied in language education context for the first time to Japanese native listeners, who were asked to reproduce what they heard in the Japanese utterances spoken by Vietnamese native speakers. It was a novel, inexpensive, and pedagogically-valid method of while-listening observation of listeners' perception. Recently in Lin, Takashima, Saito, Minematsu, and Nakanishi (2000); and in Zhu, Hakoda, Saito, Minematsu, Nakanishi, and Nishimura (2021), English RS utterances of listeners with different language backgrounds were analyzed in terms of their acoustic shadowability scores. The scores were compared not to subjective scores but to objective scores of word-level intelligibility, which was calculated by manually transcribing the RS utterances. Results showed a high correlation between the manually-calculated word-level intelligibility score of an L2 utterance and the automatically-calculated acoustic shadowability score, for any listener with any language background. This indicated a possibility of automatically predicting the word-level L2 utterance intelligibility only by asking the listeners to reverse-shadow the L2 utterance, without asking for manual transcripts.

In the current study, the data obtained in Zhu, Hakoda, Saito, Minematsu, Nakanishi, and Nishimura (2021) will be used for a different purpose. The reverse-shadowed L2 English utterances (human *while-listening* reproduction) will be compared with the transcriptions automatically generated from ASR in phoneme-level accuracy as well as word-level accuracy. Moreover, since Minematsu and Nakanishi (2021); and Zhu, Lin, Minematsu, and Nakanishi (2020) revealed that the shadowability scores vary depending on the native language of the shadowers, comparisons will be made among shadowers with different language backgrounds.

## 2. Method

### 2.1 Participants

Two groups of participants were involved in the reverse-shadowing activity; a group of 12 university students learning English in Japan, and a group of six international users of English. Upon participating in this study, they all agreed that their recorded audio would be analyzed anonymously and the results would be published for research purposes.

The former, learner group, were freshmen or sophomores ($n$ = 12) taking English conversation courses in Japan. Their English speaking proficiency was in the range of CEFR A 1 and A 2, measured by Versant English Speaking Test ($M$ = 36.0, $sd$ = 7.6). As a part of their course work, they prepared manuscripts for oral presentations and practiced reading them aloud, which were recorded and submitted through an LMS (Learning Management System). The recordings were done individually at home in late May and early June, 2020. For later reference, two English native speakers (British and American) were also asked to provide their recordings, reading short passages from a textbook aloud. These two recordings were considered to represent the utterances of quisi-learners who are very fluent in English.

The latter group, shadower group, consisted of three sub-groups, i.e., two native speakers of Japanese (NJ), two native speakers of English (NE), and two native speakers of other languages (NO; Vietnamese and Chinese). They were all fluent in English. It was confirmed that the NEs and NOs had never had experiences of learning the Japanese language, or staying in Japan, though the degrees of exposure to Japanese-accented English such as through media and their acquaintances could vary. After receiving explanations and instructions about this study and undergoing shadowing trainings with sample passages, they accessed a web page where they listened to the learners' utterances and recorded their shadowing.

To avoid the order effect (the performance could be susceptible to familiarity with the procedure and particular accents), the 14 utterances to be shadowed were presented in a random order for each shadower. The audios to be analyzed for this study are their first attempts of shadowing for each of the 14 utterances, with no manuscript shown visually.

### 2.2 Materials

Three types of transcripts, i.e., excerpts from the learners' original manuscripts (including two textbook materials read by English native speakers); the ASR transcripts of the learners' recordings; and the RS transcripts of the shadower's recordings were used for this study.

The average length of the audio files originally submitted by 12 learners and two English native speakers was 136.6 seconds ($sd$ = 66.5). Out of these files, around 30-second speech segments were manually extracted and used for the experiment. The criteria for extracting the 30-second utterances were as follows: First, the contents of the 14 utterances should not overlap, because the shadowers would easily guess what is said when the same content is repeated. Secondly, the utterances should not include proper nouns such as personal names, which would be too hard to be perceived even in authentic conversation settings. The Automatic Readability Index (ARI) of each extracted text was also calculated using Word Level Checker (Someya, 1998), to ensure that the difficulty of manuscripts did not deviate from each other significantly ($M$ = 7.3, $sd$ = 2.5). These 14 manuscripts were used as reference data as what the learners intended to say.

As for the ASR transcripts, Google API (Natal, Shires, & Jägenstedt, 2019) was used with the default settings (tuned to American English) to transcribe the 14 utterances. To confirm that the speech recognition was not affected by external factors such as background noise and microphone settings, the number of words in the ASR transcripts was compared with those in the original manuscript. The rate of the ASR transcripted word counts to the original manuscript in this study was satisfactory ($M$ = 105.8%, $sd$ = 10.3%).

For transcribing the RS recordings, three transcribers were hired. All three transcribers were fluent users of English, with different L1s. Considering that their manual transcripts may be affected by the language accents of the shadowers, a native speaker of Japanese was asked to transcribe the shadowed utterances of NJs, a native speaker of English for NEs and NO (Vietnamese, whose accent was American), and a native speaker of Chinese for NO (Chinese). They were instructed to transcribe the RS recordings word by word, carefully listening to the recordings without guessing from the context. Any unintelligible speech segment was annotated as "*", and when two unintelligible segments were found consecutively, they were annotated as "* *".

### 2.3 Calculating similarity between ASR and RS transcripts

We had two kinds of transcripts obtained for learners' utterances. One is from machines, that is ASR transcripts and the other is from human listeners, that is RS transcripts. Similarities of the two transcripts were quantified by measuring their Levenshtein distances from the learners' original manuscripts. Here, the algorithm of dynamic time warping (DTW) was applied to determine the optimal alignment between each of the ASR and RS transcripts and its original manuscript. By

treating the original manuscript as ground truth, the number of substituted words (S), deleted words (D), and inserted words (I) in the ASR and RS transcripts were calculated. By using these figures, similarities and differences of the two transcripts were examined. Further, the accuracy of ASR and RS was calculated in the following equation.

accuracy = (N−D−I−S)/N,

where N is the number of words in the ground truth, and D, I, and S are the numbers of deleted words, inserted words, and substituted words, respectively.

After comparing the word-level transcripts between ASR and RS, phoneme-level comparisons were taken as another kind of measurement. The ASR and RS transcripts were converted into their phonemic transcriptions using the CMU pronunciation dictionary (Carnegie Mellon University, 2000), which is widely used in the American English ASR community. The dictionary has 39 phonemes (15 vowels and 24 consonants) and, to each vowel of any word entry in the dictionary, lexical stresses (0: unstressed, 1: primary, 2: secondary) are also applied. Since conversion from a word to its phonemes was done solely based on the dictionary, the obtained phoneme sequences were canonical pronunciations, not actually observed phonemes in learners' utterances. With these phoneme sequences, phoneme-level deletions, insertions, and substitutions were used for calculating phoneme-level accuracy. Hereinafter, word-level and phoneme-level accuracies are referred to as WAcc and PAcc, respectively.

In word-level similarity calculation, a word pair of *walk* and *walked,* for example, are judged different (i.e., the whole word *walk* substituted with *walked*). On the other hand, in phoneme-level calculation, the difference is only one phoneme. Figure 1 shows examples of word-level (upper table) and phoneme-level (lower table) comparisons between the original manuscript, the ASR transcript, and the RS transcript of O2 (one of the NO shadowers). The notes below each of the tables show the ratios of D, I, and S.

Figure 1

*Examples of Deletion, Insertion and Substitution*

Word-level comparisons

| Original | finishing | | | | your | own | assignment | is | important |
|---|---|---|---|---|---|---|---|---|---|
| ASR | can | you | sing | to | your | own | assignment | is | important |
| O2 | finishing | | | | | * | * | silence | important |

ASR: D = 0%, I = 50%, S = 16.6%
O2: D = 16.6%, I = 0%, S = 50%

Phoneme-level comparisons

| Original | f | ɪ́ | n | ɪ | ʃ | ɪ | | ŋ | | j | ər | óʊ | n | ə | s | áɪ | n | m | ə | n | t | i̱ | z | ɪ | m | p | ɔ́:r | t | ə | n | t |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ASR | k | ə | n | j | ʊ | s | ɪ́ | ŋ | tə | j | ər | óʊ | n | ə | s | áɪ | n | m | ə | n | t | i̱ | z | ɪ | m | p | ɔ́:r | t | ə | n | t |
| O2 | f | ɪ́ | n | ɪ | ʃ | ɪ | | ŋ | | | | | * | * | s | áɪ | | l | ə | n | | | s | ɪ | m | p | ɔ́:r | t | ə | n | t |

ASR: D = 0%, I = 6.8%, S = 17.2%
O2: D = 20.1%, I = 0%, S = 13.8%

In word-level, the original manuscript is "finishing your own assignment is important…." The figure shows that the L2 vocalization of "finishing" was recognized by ASR as "can you sing to". Because the alignment program attempts to align the manuscript and the ASR transcript word by word, it claims that "finishing" was replaced by "can", and the three words of "you sing to" were judged to be inserted. Meanwhile, the O2 shadower could not identify many words in the L2 utterance. In his/her shadowing, only the first and last words were correctly shadowed, and two unintelligible speech segments were generated by the transcriber.

In phoneme-level, the O2 shadower perceived "silence" in the L2 vocalization of "assignment is". In the word-based alignment, "silence" in O2's transcript is a word substituted for "is" in the original manuscript and is counted as one substitution error. However, in the phoneme-based alignment, a few phonemes in "silence" are judged to be correctly aligned to their corresponding phonemes in the original manuscript.

Generally speaking, similarity between a manuscript and a transcript is higher in their phoneme-based alignment. We're aware that neither of word-level and phoneme-based alignments are perfect because of ambiguity problems, but we consider that statistical comparison between manuscripts and transcripts still show some indications on how we should use ASR for teaching languages.

## 3. Results and Discussions

### 3.1 Word-level and Phoneme-level Accuracies in RS and ASR

The mean values of WAcc and PAcc of the shadowed British and American English utterances

for all the six shadowers ($n$ = 12) were 95.1% ($SD$ = 4.9) for WAcc and 95.8% ($SD$ = 5.2) for PAcc, and those of the ASR transcripts ($n$ = 2) were 95.5% ($SD$ = 0.6) for WAcc and 98.1% ($SD$ = 0.2) for PAcc. That is, both the human shadowers and ASR were decently capable of recognizing British and American English utterances.

First, to grasp the overall image of the original manuscript and the corresponding transcripts, the number of words and phonemes in each text are summarized in Table 1. The shadower O2 reproduced relatively less words ($M$ = 49.8, $SD$ = 11.6) and phonemes ($M$ = 192.6, $SD$ = 46.6) than other shadowers. However, O2's WAcc and PAcc rates for the native English utterances were 96.6% ($SD$ = 3.4) and 97.0% ($SD$ = 3.0) respectively, which confirmed that O2 had an adequate ability to shadow native English utterances.

Table 1

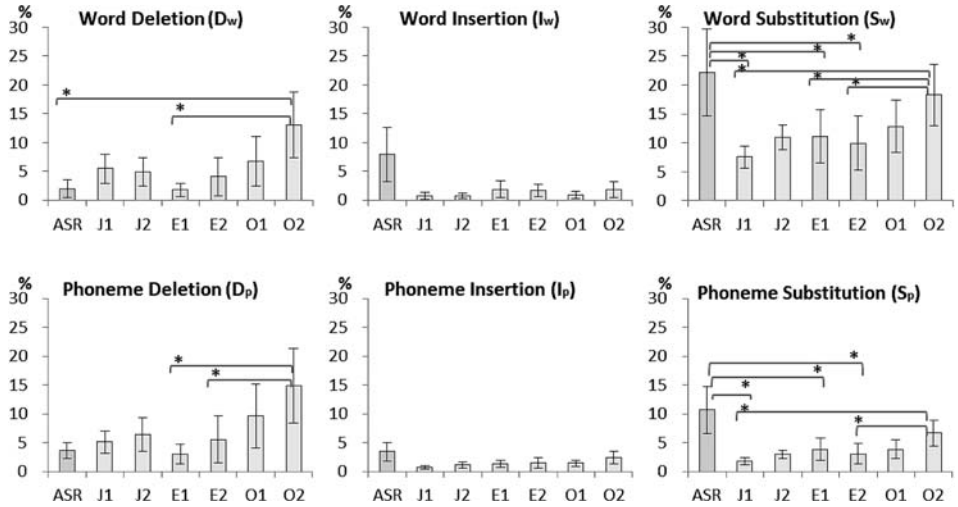
*The Mean Values and Standard Deviations of Word Count and Phoneme Count in Original Manuscripts, ASR and RS Transcripts (n = 14)*

| Counts | Original | ASR | RS | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | J1 | J2 | E1 | E2 | O1 | O2 |
| Word | | | | | | | | |
| *M* | 56.7 | 59.7 | 53.7 | 54.1 | 56.6 | 55.3 | 52.9 | 49.8 |
| *SD* | 13.0 | 12.8 | 11.2 | 12.0 | 12.3 | 13.1 | 11.8 | 11.6 |
| Phoneme | | | | | | | | |
| *M* | 222.1 | 221.3 | 211.5 | 210.3 | 218.0 | 213.2 | 202.8 | 192.6 |
| *SD* | 48.2 | 46.2 | 42.9 | 46.0 | 47.4 | 50.6 | 46.4 | 46.6 |

Second, the word-level and phoneme-level occurrences of deletion ($D_w$), insertion ($I_w$), and substitution ($S_w$), and phoneme-level occurrences of deletion ($D_p$), insertion ($I_p$), and substitution ($S_p$) in each of the seven transcripts were counted (Figure 2).

Figure 2

*Deletion, Insertion and Substitution Rates by ASR and RS*




*Note.* Upper row = Word-level counts; Lower row = Phoneme-level counts.
*$p < .05$. Error bars show 95% CI.

To examine the variability of the D-, I-, and S- rates among the seven transcripts, One-way ANOVAs (7 levels: ASR, J1, J2, E1, E2, O1, O2) were conducted for word-level and phoneme-level D, I, and S rates. Significant differences were found in all of the analyses ($p < .001$ for $D_w$, $S_w$, and $S_p$; $p < .01$ for $D_p$ and $I_p$, and $p < .05$ for $I_w$). Further, multiple comparisons for the transcripts were conducted, adjusted with Holm's sequentially rejective Bonferroni procedure. Statistically significant differences were found between the pairs indicated with * in Figure 2.

Briefly observing Figure 2, the $D_w$ rates seem to be similar to the $D_p$ rates. As described earlier in the previous section, when a word is deleted, all the constituent phonemes in the word are also deleted and thus, relative differences between $D_w$ and $D_p$ are small. The $I_w$ rate for ASR is higher than its $I_p$ rate, which indicates that the ASR system has a tendency of inserting a word sequence that sounds similar to its corresponding word in the original manuscript. The $S_p$ rates are smaller than the $S_w$ rates across ASR and RSs, indicating that the misrecognized words tend to have some phonemes which correspond to the ones in the original manuscripts.

### 3.2 Variability within Shadowers

Among the six shadowers, statistically significant differences were found between the shadower O2 and others in $D_w$: $t(13) = 4.50$, $p = .013$, $r = .78$ with E1; in $S_w$: $t(13) = 4.75$, $p = .007$, $r = .80$ with J1; $t(13) = 4.16$, $p = .018$, $r = .76$ with E1; $t(13) = 5.28$, $p = .003$, $r = .83$ with E2; in $D_p$: $t(13) = 4.24$, $p = .020$, $r = .76$ with E1; $t(13) = 3.86$, $p = .039$, $r = .73$ with E2; in $S_p$: $t(13) = 4.32$, $p = .$

016, $r = .77$ with J1; $t(13) = 3.74$, $p = .042$, $r = .72$ with E2.

As described in section 2.1, the six shadowers can be divided into three sub-groups, i.e., J1 and J2 who are accustomed to Japanese accented English; E1 and E2 who are native speakers of English; and O1 and O2 who are native speakers of Vietnamese and Chinese respectively. The result in the current study supported the findings in Minematsu and Nakanishi (2021); and Zhu, Lin, Minematsu, and Nakanishi (2020), in that L2 learners' accented speech is not always intelligible to non-native speakers of English who do not share the same L1 with the learner. For example, the $S_w$ rate of the shadower O2 was significantly larger than that of J1, E1, and E2. That is, the shadower O2 was more likely to misperceive the learners' Japanese accented utterances. If a learner were to communicate with an international English user like O2, the risk of miscommunication is higher than in communication with native speakers of Japanese or of English.

In Japanese classroom settings, English teachers are often native speakers of Japanese or English. Thus, from a learners' point of view, it should be kept in mind that being understood by a teacher in class does not necessarily mean that their utterances are intelligible enough in an international communication setting. At the same time, from a teachers' point of view, ways for spotting learners' unintelligible pronunciation need to be considered. The teachers whose L1 is the same as the learners' target language, and those who share the same L1 with the learners may not be able to realize the level of intelligibility in authentic communication with users of English with various language backgrounds.

### 3.3 Comparisons between Human Shadowers and ASR

From the comparisons between RS and ASR, statistically significant differences were found between ASR and shadowers in $D_w$: $t(13) = 3.76$, $p = .048$, $r = .72$ with O2; in $S_w$: $t(13) = 4.22$, $p = .017$, $r = .76$ with J1; $t(13) = 4.25$, $p = .017$, $r = .76$ with E1; $t(13) = 5.17$, $p = .004$, $r = .82$ with E2; in $S_p$: $t(13) = 4.23$, $p = .018$, $r = .76$ with J1; $t(13) = 4.39$, $p = .015$, $r = .77$ with E1; $t(13) = 4.49$, $p = .013$, $r = .78$ with E2.

All the ASR and O2 transcripts were reviewed and qualitatively compared with the original manuscripts to examine how different the instances of ASR are from human RS. A close examination of these texts revealed three characteristics of ASR: unrealistic choice of words, stylistic inconsistencies, and substitution with technical jargon. Figures 3 to 5 are the excerpts of the original manuscripts and the transcripts by ASR and O2 transcripts, where these characteristics are reflected.

As can be seen in figure 3, "working generation" is transcribed as "walking gelation" by ASR. From a phonetic point of view, this is a reasonable transcription, in that Japanese accented speakers often confuse /ɚ́:r/ in *work* with /ɔ́:/, and in that /r/ is often pronounced interchangeably with /l/.

Indeed, *working* was pronounced more like *walking* and the consonant /r/ in *generation* was pronounced closer to /l/ in the learners' original recording. However, prior to this phrase, the learner explained about the problem of ageing society and lack of nursing homes. This is one of the features of ASR mentioned in Benzeghiba, et. al. (2007), being weak at representing grammatical and semantic knowledge. It is partly because ASR generally recognizes the input utterances phrase by phrase, independently from each other, i.e., the contents of the utterance are not interpreted beyond the boundaries of sentences. On the other hand, O2 was able to shadow the phrase "the working generation" as the learner intended. Though the shadowers were asked to avoid guessing the utterance from the context, it is natural that humans try to make sense of the words they hear as much as possible.

Figure 3

*Excerpt #5 from Learners' Original Manuscripts and Transcripts by ASR and RS*

| Original | the | working | generation | have | to | give | more | support | for | elderly |
|---|---|---|---|---|---|---|---|---|---|---|
| ASR | the | walking | gelation | how | to | give | more | sports | for | elderly |
| O2 | the | working | generation | are | * | | more | support | for | elderly |

In excerpt #22 (figure 4), the sentence generated by ASR starts with "yeah". This colloquial style is partly due to the default setting of the ASR system. When using ASR for pedagogical purposes, it is necessary for the users to consider the preferred language style suitable for language learning. On the other hand, the shadowed utterance by O2 was full of deletions and substitutions. It is especially notable that the word "syllabus" was shadowed as "problem", which does not contain similar phonemes to the intended word. This phrase was uttered in the middle of the 30-second recording, following an explanation of how their academic achievement is graded, including a phrase "whether you passed or failed". It is possible that the word "fail" was kept in O2's mind, which caused the misunderstanding that the learner was talking about some kind of problems. This mis-shadowed utterance suggests the reality of human communication, where unexpected misunderstandings can derive from word association rather than mispronunciation.

Figure 4

*Excerpt #22 from Learners' Original Manuscripts and Transcripts by ASR and RS*

| Original | they | are | assessed | based | on | | the | grading | policy | on | the | syllabus | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ASR | yeah | assess | the | ways | to | earn | the | grading | policy | on | the | shuttle | bus |
| O2 | there | are | access | based | | | | | rating | policies | additional | problems | |

Finally, the last word "actually" in excerpt #25 was recognized as "actuary", which is a word not

likely to be used by a learner who is not fluent enough in English, unless he or she is a specialist in the field. This indicates a problem of using ASR for English learning without much consideration. As described in section 1.1, ASR systems were not originally developed for L2 learners. Because the ASR transcript is dependent on the acoustic and language models (Cucchiarini & Strik, 2018), when the models are based on the L1 users' corpus, the language output may not be suitable for describing language learners' fluency. On the other hand, the word "actuary" did not seem to occur in O2's mind when he shadowed the speech, even though the learner's pronunciation was closer to "actuary" than "actually". This incident suggests the nature of human communication, where the speaker and listener try to adjust their language by choosing words that are appropriate for each other's language fluency.

Figure 5

*Excerpt #25 from Learners' Original Manuscripts and Transcripts by ASR and RS*

| Original | they | were | getting | ready | for | Valentine's | | Day | sales, | actually |
|---|---|---|---|---|---|---|---|---|---|---|
| ASR | they | robbed | getting | Daddy | for | barren | tines | day | says | actuary |
| O2 | they | are | getting | ready | for | valentine's | | day | sales | actually |

**4. Conclusion and Future Research**

In this study, English learners in Japan ($n$=12) and native speakers of English ($n$=2) read aloud their speech manuscripts, recordings of which were reverse-shadowed by three groups of shadowers with different language backgrounds. The transcripts of these shadowed utterances were compared with the ASR transcripts of the same recordings, to examine how differently L2 learners' intended speech is recognized by RS and ASR. Their word-level and phoneme-level deletion, insertion, and substitution rates were calculated by referring to the learners' manuscripts. The results of the comparison indicated that shadowability of L2 speech can be affected by the language background of shadowers, and ASR transcription does not necessarily simulate human perception well.

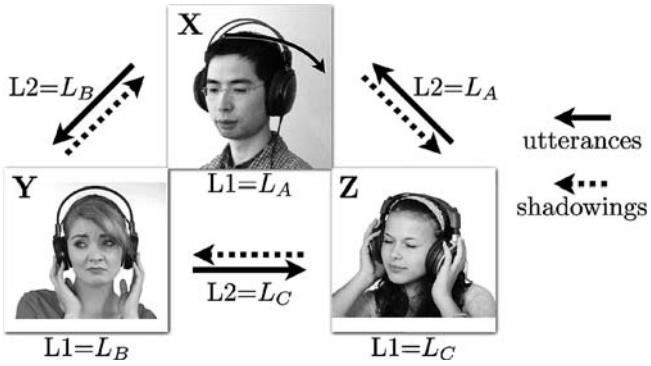
Following the results, we consider that adequately good and adequately poor systems are needed for assessment, which have a human-like robustness to foreign accents. To realize such systems, it is evident that both goodness and poorness in human performance of transcription should be modeled. If an L2 utterance is given to a human transcriber, some parts will be transcribed correctly and others will not. If teachers want a machine that can simulate human transcribers, a huge paired collection of L2 utterances and human transcriptions, which include many errors, have to be prepared. For a given utterance, its manuscript is unique, its transcripts are diverse depending on the language backgrounds of transcribers. Considering this fact, a huge enough collection of L2 utterances and diverse enough transcripts is infeasible. We consider that reverse shadowing is a

possible solution, which can be viewed as instantaneous and oral transcription. In Zhu, Lin, Minematsu, and Nakanishi (2020), acoustically-derived shadowability scores were found to be highly correlated to intelligibility scores derived from manual transcription. Although human shadowers were used to calculate acoustic shadowability scores in this study, the first attempt to develop a machine shadower was made in Zhu (2021). After collecting a large enough paired collection of L2 uttearnces and RS utterances, we will build machine shadowers, which will be able to give shadowability scores as feedback to English learners without using human shadowers.

To realize a huge enough collection of L2 utterances and RS utterances, inter-learner shadowing (ILS) may be a pedagogically-sound solution, conceptually illustrated in Figure 4. Learner X, who speaks Language A as L1 and is learning Language B, reads aloud sentences in Language B. His utterances are shadowed by learner Y, who speaks Language B as L1 and is learning Language C. Her utterances are shadowed by learner Z, who speaks Language C as L1 and is learning Language A. Her utterances are shadowed by learner X. ILS is language learning and language exchange activity, where every learner is supporting others as native speaker and is supported by others as language learner.

Figure 4

*Inter-learner Shadowing*




**Acknowledgements**

This work was supported by JSPS KAKENHI Grant Number 17K02914, 18H04107, and 20K20407.

**References**

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., & Wellekens, C. (2007). Automatic speech recognition and speech variability A review. *Speech communication, 49*(10-11), 763-786. https://doi.org/bwvd39

Carnegie Mellon University. (2000). The CMU pronunciation dictionary. http://www.speech.cs.cmu.edu

Coniam, D. (1999). Voice recognition software accuracy with second language speakers of English. *System, 27,* 49-64. https://doi.org/b3k97n

Cucchiarini, C. & Strik, H. (2018). Automatic speech recognition for second language pronunciation training. In O. Kang, R. I. Thomson, & J. M. Murphy (Eds.), *The Routledge Handbook of Contemporary English Pronunciation* (pp.556-569). Routledge.

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence for four L1s. *Studies in Second Language Acquisition, 20,* 1-16. https://doi.org/bjstvw

Derwing, T. M., Munro, M. J., & Carbonaro, M. D. (2000). Does popular speech recognition software work with ESL speech? *TESOL quarterly, 34*(3), 592-603. https://doi.org/fwphxx

Goslin, J., Duffy, H., & Floccia, C. (2012). An ERP investigation of regional and foreign accent processing. *Brain and Language, 122*(2), 92-102.

Govender, A., & King, S. (2018). Using pupillometry to measure the cognitive load of synthetic speech. In *Proceeding of INTERSPEECH 2018,* 2838-2842.

Hincks, R. (2002, September 16-20). Speech recognition for language teaching and evaluating: A study of existing commercial products. In *Proceeding of Seventh International Conference on Spoken Language.* https://rb.gy/kidgpa

Inoue, Y., Kabashima, S., Saito, D., Minematsu, N., Kanamura, K., & Y. Yamauchi. (2018). A study of objective measurement of comprehensibility through native speakers shadowing of learners' utterances. In *Proceeding of INTERSPEECH 2018,* 1651-1655.

Jenkins, J. (1998). Which pronunciation norms and models for English as an International Language? *ELT Journal, 52*(2), 119-126. https://doi.org/dt6mnj

Jenkins, J. (2005). Implementing an international approach to English pronunciation: The role of teacher attitudes and identity. *TESOL quarterly, 39*(3), 535-543. https://doi.org/fhwmxn

Kadota, S. (2019). *Shadowing as a practice in second language acquisition: Connecting inputs and outputs*. Routledge.

Levis, J. (2005). Changing contexts and shifting paradigms in pronunciation teaching. TESOL *Quarterly, 39,* 369-77. https://doi.org/d5dbvn

Levis, J. (2020). Revisiting the intelligibility and nativeness principles. *Journal of Second Language Pronunciation. 6*(3), 310-328. https://doi.org/fj9f

Lin, Z., Takashima, R., Saito, D., Minematsu, N., & Nakanishi, N. (2000). Shadowability annotation with fine granularity on L2 utterances and its improvement with native listeners' script-shadowing. In *Proceeding of INTERSPEECH 2020,* 3865-3869.

McCrocklin, S. (2016). Pronunciation learner autonomy: The potential of automatic speech recognition. *System, 57,* 25-42. https://doi.org/f8fmr7

McCrocklin, S. (2019). ASR-based dictation practice for second language pronunciation improvement. *Journal of Second Language Pronunciation, 5*(1), 98-118. https://doi.org/fj9g

Minematsu, N., & Nakanishi, N. (2021). Automatic calculation of instantaneous intelligibility based on reverse shadowing: How smoothly are L2 utterances understood by listeners?. In *Proceeding of the 47th National Conference,* Japanese Association for Asian Englishes (JAFAE).

Mroz, A. (2018). Seeing how people hear you: French learners experiencing intelligibility through automatic speech recognition. *Foreign Language Annals, 1*(21), 1-21. https://doi.org/fj9h

Munro, M. (2008). Foreign accent and speech intelligibility. In J. G. H. Edwards and M. L. Zampini (Eds.), *L2*

*speech production research: Findings, issues and advances* (pp. 193-218), John Benjamins.

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 45*(1), 73-97. https://doi.org/cmfdjv

Nakanishi, N., Tam, S. Y., & Ebihara, Y. (2020). Spotting English pronunciation errors: Comparison among teachers and with automatic speech recognition. *LET Kansai Chapter Collected Papers, 18,* 125-146.

Natal, A., Shires, G., & Jägenstedt, P. (2019). Web Speech API: Draft Community Group report, September 30, 2019. https://wicg.github.io/speech-api/

Neri, A., Cucchiarini, C., & Strik, H. (2003). Automatic speech recognition for second language learning: How and why it actually works. In *Proceeding of International Congresses of Phonetic Sciences,* 1157-1160. https://rb.gy/xypcet

Romero-Rivas, C., Martin, C. D., & Costa, A. (2016). Foreign-accented speech modulates linguistic anticipatory processes. *Neuropsychologia, 85,* 245-255.

Shadiev, R., Hwang, W. Y., Chen, N. S., & Huang, Y. M. (2014). Review of speech-to-text recognition technology for enhancing learning. *Journal of Educational Technology & Society, 17*(4), 65-84. https://rb.gy/mz5wvv

Someya, Y. (1998). Word Level Checker. [online software]. http://someya-net.com/wlc/

Zhu, C. (2021). Sequential annotation and prediction of L 2 speech instantaneous intelligibility based on shadowing techniques. Master thesis of Graduate School of Engineering, The University of Tokyo.

Zhu, C., Hakoda, R., Saito, D., Minematsu, N., Nakanishi, N., & Nishimura, T. (2021). Multi-granularity annotation of instantaneous intelligibility of learners' utterances based on shadowing techniques. In *Proceeding of IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop,* 1071-1078.

Zhu, C., Lin, Z., Minematsu, N., & Nakanishi, N. (2020). Analyses on instantaneous perception of Japanese English by listeners with various language profiles. In *Proceeding of the 34th General Meeting,* 26-31, Phonetic Society of Japan (PSJ ).