The Role of ASR Training in EFL Pronunciation Improvement: An In-depth Look at the Impact of Treatment Length and Guided Practice on Specific Pronunciation Points

Ryan Spring (spring.ryan.edward.c4@tohoku.ac.jp)
Tohoku University, Japan

Ryuji Tabuchi (tabuchiryuji@nifty.ne.jp)
Mint Phonetics Education Institution, Japan

Abstract

Several studies have suggested that pronunciation practice through automatic speech recognition (ASR) tools can help learners improve second language pronunciation. However, the treatment length varies from study to study, making it unclear whether a longer treatment at shorter intervals or a more intensive treatment for longer intervals will have a greater impact. Furthermore, since most studies include both teacher instruction and ASR-based practice, it is unclear how much impact is due to the feedback and guided practice of the ASR tools and how much is due to teacher instruction. This study seeks to discover if ASR-based practice has a measurable impact on student performance discernable from teacher instruction, which pronunciation points are most impacted by such practice, and whether treatment length affects learning outcomes. We found that L1 Japanese English as foreign language (EFL) learners were more likely to improve on vowel-related pronunciation and that while treatment length over a single semester did not have a large impact on learning outcomes, the feedback from an ASR-based practice tool caused students to focus on their pronunciation and adjust it, often for the better. Therefore, we conclude that ASR-based pronunciation tools are meaningful in a wide variety of Japanese EFL contexts and recommend their usage.

Keywords: Automatic Speech Recognition, Pronunciation, English as a Foreign Language, Feedback

Introduction

The use of automatic speech recognition (ASR) in the foreign language classroom to help learners improve their second language (L2) pronunciation is recently gaining increasing attention (e.g., Author, 2021; Golonka et al., 2014; Mroz, 2018; Xiao & Park, 2021). However, there is some disagreement in the literature regarding how much impact ASR-based activities have on learners' pronunciation and how to utilize ASR technology as part of the learning process. Studies such as Bozorgian and Shamsi (2020), Guskaroska (2019), and Spring and Tabuchi (2021) found that the use of ASR helped learners' overall pronunciation, but they often observed different pronunciation points. For example, Bozorgian and Shamsi (2020) focused on suprasegmental pronunciation, and Guskaroska (2019) focused on vowel pronunciation. Spring and Tabuchi (2021) attempted to observe the impact on several different pronunciation points. Their findings showed that ASR-based training had an overall positive impact on pronunciation ability but that the effects varied from learner to learner, leading them to suggest a more in-depth analysis of individual pronunciation points as well as a study that was more long-term than the 5-week treatment they reported on. Furthermore, as most studies employ pre- and post-tests to show improvement but include instruction as well as ASR training, the impact of the feedback provided by ASR in the acquisition processes is still unclear. The purpose of this study is to improve on previous studies by (i) offering new insights into the impacts of ASR-based pronunciation practice over an extended period of time and (ii) elucidating the impact of retrying drills on specific pronunciation points.

Literature Review

Intelligibility and ASR

Learning pronunciation in an L2 is essential because it is the foundation of speaking, and without a certain degree of pronunciation accuracy, others will be unable to understand the speakers' utterances. However, there has been considerable debate about what areas of pronunciation are most important, which should be focused on in the L2 classroom, and what is an acceptable level of pronunciation mastery for learners. While there are a number of areas in which pronunciation can be evaluated, such as comprehensibility, interpretability, and accentedness, recent studies have suggested that these points are often flawed due to human perception error (e.g., Hu & Su, 2015; Lindemann & Subtirelu, 2013). Therefore, many studies suggest that intelligibility, i.e., how well the learner can be understood by others, can be used as an acceptable evaluation criterion and that being intelligible is a reasonable goal for L2 speakers to strive for (e.g., Levis, 2018; Munro, 2010).

Measuring intelligibility is another challenge for L2 teachers and learners. Though some studies have employed native speakers and asked them to transcribe L2 learners' speech as a metric, this process can be flawed due to human error (e.g., Lindemann & Subtirelu, 2013) and is also both costly and time-consuming, which makes it impossible to implement practically in L2 classroom settings. Here, many studies have begun to employ ASR-based tools because they are objective and perfectly reliable (i.e., they will produce the same scores for the same sound every time), but also free or low-cost and require almost no time to deliver results. Furthermore, several studies have suggested that the more accurately ASR can transcribe an L2 learners' speech, the higher native speaking judges tend to rank them in terms of pronunciation ability (e.g., Author, 2020; Ashwell & Elam, 2017; Guskaroska, 2019; Mroz, 2018). For these reasons, ASR-based tools have become more and more commonplace in both the teaching and evaluation of L2 pronunciation and general speaking. However, there are still discrepancies in how these tools are best used in the teaching and L2 learning process.

ASR-based EFL Pronunciation Practice

Several studies have been conducted on the use of ASR-based pronunciation practice in EFL teaching (e.g., Ahn & Lee, 2016; Bozorgian & Shamsi, 2020; Evans & Chen, 2020; Guskaroska, 2019; 2020; Inceoglu et al., 2020; McCrocklin, 2019; Xiao & Park, 2021). Many of the studies are focused on student and teacher experiences with ASR-based tools, generally finding that the tools are received positively, albeit to varying degrees of satisfaction, depending on how the tools were integrated into the learning process (e.g., Ahn & Lee, 2016; Hsu, 2015; Sidgi & Shaari, 2017; Wang & Crosthwaite, 2021; Xiao & Park, 2021). Other studies provide case studies of how ASR tools can be useful for EFL learners in their context (e.g., Author, 2021; Bozorgian & Shamsi, 2020; Guskaroska, 2019; McCrocklin, 2019). For example, Bozorgian and Shamsi (2020) demonstrated that using ASR for several sessions over a two-month period could help improve the pronunciation of suprasegmental features in lower-level EFL learners. Furthermore, Xiao and Park (2021) found that ASR technology enhanced the learning of learners with a wide range of individual differences. Other studies have attempted to provide broader data sets and focus on particular pronunciation points. For example, Guskaroska (2019) found empirical evidence that advanced Macedonian EFL learners improved at specific vowel sounds and pronunciation in general. Author (2021) observed the impact of ASR training on Japanese EFL learners and similarly found that it helped improve learners' overall pronunciation but that the effects were much greater for lower-level learners. Finally, reviews such as Glonka et al. (2014) have determined that ASR and other computer-assisted pronunciation training seem to generally be useful for EFL learners.

However, there are still areas that require more research on this topic. One area that should be explored is the length of the treatment. Guskaroska (2019), Spring and Tabuchi (2021), and Bozorgian and Shamsi (2020) reported on focused treatments over short terms (two, five, and eight weeks, respectively), but it is still unclear how students will be affected by treatments that are individually shorter but spread over a longer term. As pronunciation is a productive skill that requires practice, longer treatment times may lead to greater improvement, but this is still difficult to verify due to the somewhat low comparability between previous studies, i.e., there are significant differences in length of treatments, L1s, and beginning L2 proficiency. Since both L1-L2 pairings and L2 proficiency have been suggested to impact the effectiveness of pronunciation training (e.g., Chau et al., 2022; Glonka et al., 2014; Spring & Tabuchi, 2021), a comparative study that changes only treatment times could provide clearer insight into how this single factor affects ASR-based training. Furthermore, there is still uncertainty with regards to which specific points ASR can and cannot help within specific contexts. For example, Inceoglu et al. (2020) found that only certain segmental aspects of L1 Korean learners' EFL pronunciation were improved through ASR training. Moreover, though Guskaroska (2019) and Bozorgian and Shamsi (2020) reported improvement in specific pronunciation points (vowel sounds and suprasegmental features, respectively), Spring and Tabuchi (2021) did not observe an effect on any single specific pronunciation points as compared to others. Rather, Spring and Tabuchi (2021) noted a large amount of variation in individual learners. Therefore, a more in-depth study of specific consonant sounds and other errors and how each individual reacted to the ASR training could provide more insight into how it impacts particular pronunciation points for particular types of learners.

Feedback and Self-Correction in Computer-assisted Pronunciation Training

Feedback is generally thought to be important to learning in general, including EFL education, but especially so in EFL pronunciation training (e.g., Wang & Young, 2012; 2015). In general, feedback for computer-assisted pronunciation training has been offered in three modes: textual, audio, and visual (Wang & Young, 2015). Visual feedback is often provided as waveforms produced by the learners and by native speakers for comparative purposes (e.g., Mehrpour et al., 2017); audio feedback is usually given as a recording of a native speaker or a text-to-speech reading when pronunciation mistakes are made (e.g., Neri et al., 2008), and text feedback is generally given as either explicit instruction or a representation of what an ASR system guessed as a transcription (e.g., Spring & Tabuchi, 2021). While various forms of feedback can be useful, Wang and Young (2015) surveyed several learners using an ASR-based tool. They found that, in general, learners appreciated corrective feedback, particularly in textual and audio modes. They compared their results to those of Chen (2011), Kim and Kim (2012), and Neri et al. (2008), and concluded that text-based description that clearly indicates learners' faults is a significant form of feedback. However, Wu et al. (2022) found that students improved their pronunciation more when peer feedback was added to feedback from an ASR alone. Moreover, both Evans and Chen (2020) and Dai and Wu (2021) reported that the data in their studies suggest that adding peer feedback to the learning process was more effective at improving pronunciation than individual practice with ASR feedback alone. Furthermore, some studies report that though they found high levels of correlation between human judgements of learner pronunciation and ASR, students are sometimes frustrated by the ASR and some even doubt its ability to properly judge pronunciation (e.g., Guskaroska, 2020; Inceoglu et al., 2020). Therefore, it is still unclear exactly how useful ASR-given feedback is for learners.

One area that is particularly lacking in the literature regarding the effects of ASR-provided feedback on learners is the amount of influence that it has on an individual level. Though learners may feel that various forms of feedback are helpful, and ASR-based tools show the ability to improve the general pronunciation ability of a population of learners, it still leaves the question of why some learners improve more than others (i.e., Author, 2021), and why some particular learners get frustrated or perform better when also provided with peer-based feedback (e.g., Dai & Wu, 2021; Inceoglu et al., 2020; Wu et al., 2022). For example, it would be informative to know if learners actually attempted to change their pronunciation in time-based on feedback and, if so, how much that leads to improvement in pronunciation for that individual. Therefore, an analysis of how much impact ASR tools and feedback have on learners' ability to modify and tune their pronunciation to be more appropriate upon repeated attempts (the "tuning" power of the tools) may elucidate why some learners improve more than others. This would also allow us to observe the impact of ASR training in the acquisition processes and assess how much of the long-term improvement is due to training and feedback and how much is due to instruction.

Based on the aforementioned studies, it is reasonable to believe that an ASR-based tool can help students improve their pronunciation. Still, the effects of tool usage over a longer period on specific pronunciation points are unknown. Furthermore, it is unknown how much tuning power ASR-based feedback has and what impact that has on correcting pronunciation errors. Therefore, this study seeks to answer the following research questions:

1. How does implementing ASR-based pronunciation practice over a semester impact L1 Japanese EFL students' pronunciation ability compared to a short period of practice?
2. What specific pronunciation points do and do not show improvement after L1 Japanese EFL students receive ASR-based training?
3. How much tuning power does ASR feedback have on learners, and what impact does it have on their pronunciation improvement?

Methodology

Research Design

This study employs a quasi-experimental design due to the fact that it was conducted on the students designated to the researcher's class. Furthermore, the only students whose data were included in this study were those who turned in their assignments and agreed to have their data collected. Therefore, the study does not meet the requirements of true random sampling needed to be considered truly experimental. It follows a basic pre- post-test design with classroom procedures and data collection conducted similarly to Spring and Tabuchi (2021) for maximal comparability. Further analyses were conducted on this data in order to answer the other research questions.

Participants

Data were taken from 19 Japanese EFL learners in their 2nd year of the university who participated in this study. They had been studying English for seven years but were lower level and could be considered CEFR A2 level as per their TOEIC® scores (see Table 1). The participants were taking an English communication class taught by one of the researchers. Though the class consisted of many more students, only 19 gave permission to participate in the study and provided both pre- and post-tests.

Table 1

Participant Data

Age	**TOEIC Score**	**Sex**
19-20 (*M*=19.16, *SD*=0.37)	240 – 510 (*M*=398, *SD*=74.7)	Male: 9, Female: 10

Classroom Procedure

Participants used a textbook as part of their class which focused on dictogloss activities and then provided discussion topics for the students. Instead of utilizing the discussion topics, the instructor provided pronunciation training for 15 minutes. During most pronunciation training sessions, a five-minute, one-point pronunciation lesson based on two individual sounds that can be difficult for L1 Japanese EFL learners to distinguish between (Goto, 1971; Mochizuki, 1981; Nakata & Shockey, 2011; Otake et al., 1993; Spring & Tabuchi, 2021). A list of the points taught can be found in Table 2.

Table 2

Pronunciation Lessons by Week

Week	Pronunciation Lesson
1	Introduction of tool and lesson-based practice
2	*r* and *l*
3	*v* and *b*
4	*θ* and *s*
5	*ð* and *z*
6	ɕ and ç
7	Review of consonants and lesson-based practice
8	*æ* and *a*
9	*ɪ̞* and *i*
10	*ʌ* and *ʊ̞* and *a*
11	*ä* and *o*
12	Review of vowels and lesson-based practice
13	Lesson-based practice
14	Lesson-based practice
15	N/A

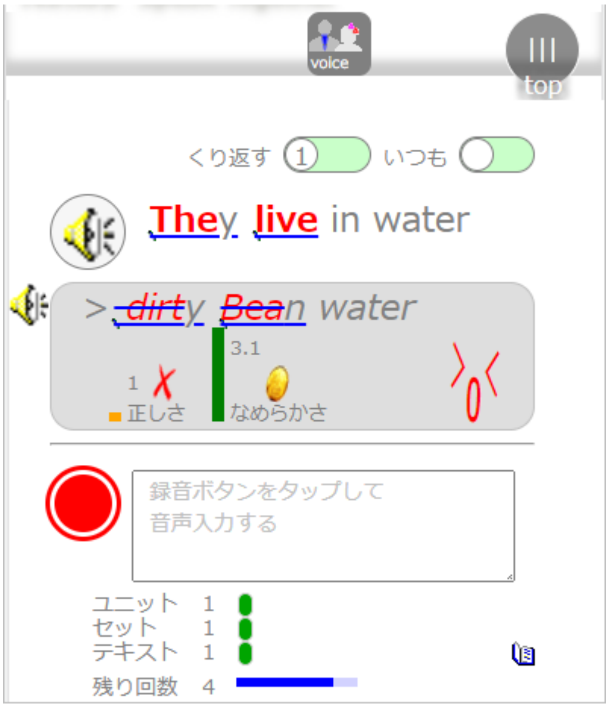
After each pronunciation lesson, students were then provided a link for practicing pronunciation using the NatTos ASR-based pronunciation tool, also utilized in Spring and

Tabuchi (2021), to ensure comparability between this study and Spring and Tabuchi (2021). The NatTos tool allows instructors to create sets of words and phrases for students to practice and a link for students. When students click on the link, they are provided with a text-to-speech listening sample of the word or phrase and can also see the word or phrase on the screen. Students can then choose to modify the text-to-speech voice, listen to the sample word or phrase again, or attempt to pronounce the word. When students choose to pronounce the word or phrase, an ASR listening instance is created, which takes input from the user's microphone and transforms the user's speech into text. The text is then displayed on the screen below the target word, or phrase, and any differences in what the ASR guessed, and the target word or phrase are highlighted. Additionally, the NatTos tool provides pronunciation and fluency scores. The pronunciation score is calculated as the percentage of letters that the ASR correctly identified in the user's speech, fit to a scale from one to ten, and rounded to one decimal place. Fluency scores are calculated as words per second based on the number of words the ASR detects the user saying and the amount of time between the user pushing the speak button and the end of the ASR detection phase. Finally, the NatTos tool offers suggestions as to what pronunciation point was incorrect that caused the error (see Figure 1). For example, if the target phrase was "get the lightbox" and the ASR guessed "get the right box," the NatTos tool will suggest that the learner should be careful with the R/L sound pairing.

If the user pronounces the target word or phrase correctly, they automatically move on to the next word or phrase in the instructor-created set. If the user does not pronounce it correctly, they are allowed another chance to listen to the target word or phrase and make another attempt at pronouncing it. Users are allowed up to five tries at each target word or phrase, but after failing for the fifth time, they are automatically transferred to the next word or phrase in the set. The data for each participant (i.e., pronunciation and fluency scores, how many times they attempted each target word or phrase and what the ASR recorded each time) are stored and available for the instructor to see.

Figure 1

Feedback Provided by the NatTos Tool upon a Pronunciation Error



They live in water

> dirty Bean water

録音ボタンをタップして
音声入力する

The pronunciation practice in this study included two sets of minimal pairs, as was done in Spring and Tabuchi (2021), but then 16 additional practice drills were provided to students which were related to the textbook lesson. Specifically, students were asked to pronounce key vocabulary words from the lesson and repeat phrases and sentences that appeared in their dictogloss listening activity. On days without a one-point pronunciation lesson, students were only provided a link for the ASR-based pronunciation tool, and all drills were focused on the textbook lesson. The instructor checked students' progress and offered specific advice when applicable based on the feedback from the tool.

Pronunciation Errors Observed in this Study

A number of pronunciation errors, more specific than those of Spring and Tabuchi (2021) but broad enough to result in no error identifying them, were selected for specific observation. First, R/L sounds (*r* and *l*), V/B sounds (*v* and *b*), and TH/S/Z (*θ* and *s*, *ð* and *z*) sounds were selected as specific consonant-based pronunciation errors to check for improvement because of the prevalence of these errors in L1 Japanese EFL learners (e.g., Goto, 1971; Mochizuki, 1981; Spring & Tabuchi, 2021). Next, pronunciation points regarding vowel sounds were divided into two categories: pronunciation of similar single vowel sounds (e.g., *ı̩* and *i*, *ʌ* and *ʊ̩* and *a*, *ä* and *o*) and epenthesis (i.e., pronouncing English words with Japanese moras causing pronunciation with more syllables than the target word). Though several individual vowel sounds were taught in the class, it was difficult for individual raters to later distinguish which individual vowel sound was pronounced incorrectly which resulted in an intelligibility error. Conversely, the problem of epenthesis is quite distinguishable and a common mistake by L1 Japanese EFL speakers (e.g., Nakata & Shockey, 2011; Otake et al., 1993) and thus included as a separate pronunciation error from individual vowel sound errors. All other errors were classified as “other” errors for the purposes of this study.

Data Collection

This study copied the pre- and post-test design of Spring and Tabuchi (2021) for checking changes in pronunciation ability to ensure comparability between the studies. Specifically, participants were asked to make a recording of themselves reading a 200-word statement that was prepared beforehand and based on a sample answer to a TOEFL iBT® test speaking question that requires an opinion to be given (e.g., Goodine, 2019; Spring & Tabuchi, 2021). Participants were asked to focus on pronunciation and not fluency to produce the most phonetically accurate recording that they could. The same script and task were used for both pre- and post-tests, but not revisited at any point during the 15-week class. Participants completed the pre-test after the first lesson and the post-test directly after the 15th lesson.

Additionally, participant data was taken from the NatTos tool. Specifically, this study considers the number of attempts and successes for target words and phrases and also uses specific examples of mistakes that participants made.

Data Analysis

The pre- and post-test data were analyzed first by identifying the number of intelligibility errors in each script. This was accomplished by first using YouTube’s ASR-based subtitling system to create a script for each recording, following Spring (2020). Though ASR transcriptions have more difficulty recognizing L2 speech than native speech, this gap has been decreasing rapidly (McCrocklin & Edalatishams, 2020), and the errors in transcription for recent ASR systems generally correlate with intelligibility errors that can lead to error pattern detection (e.g., McCrocklin, 2019b; Wallace, 2016). However, it should be noted that McCrocklin and Edalatishams (2020) show that errors in Google’s transcriptions, on which YouTube’s are based, vary depending on L1. Specifically, they found that L1 Chinese speakers’ comprehensibility and accentedness are more correlated with transcription errors than those of L1 Spanish speakers. However, Spring (2020) has shown that errors in transcripts of L1 Japanese EFL learners provided by YouTube’s ASR-based subtitles, specifically, are well correlated with errors in intelligibility. Therefore, this same system was used to detect errors for pre- and post-tests in this study. Next, the text comparison software (text-compare.com) was used to quickly identify words that YouTube’s ASR incorrectly transcribed. Two assistants then counted the total number of inconsistencies and checked each to determine and count the types of pronunciation errors (as per section 3.2) that caused the inconsistencies. The assistants were both native English-speaking graduate students who had at least four years of experience teaching English in Japan and were thus familiar with common L1 Japanese pronunciation errors that impede intelligibility. Any differences in the ASR transcript and original script that were not due to pronunciation errors, such as homonyms (e.g., “you’d” and “you would,” “their” and “there”) were not counted. Errors that were clearly due to a specific pronunciation error (e.g., “agree” appearing in a transcript as “ugly” clearly indicates an R/L pronunciation error) were logged as such, but for indeterminable discrepancies in the scripts, the assistants consulted the audio file to determine the pronunciation error type. For any disagreement between the assistants, the lead researcher was consulted, and a final decision was made (*N*=3). The differences in the total number of errors and the number of each specific pronunciation error in the pre- and post-tests were compared using dependent t-tests, with Cohen’s d used as a measure of effect size. Cohen’s d was interpreted according to the definitions of largeness provided by Plonsky and Oswald (2014). These measures were chosen both due to the continuous and normal nature of the data and for direct comparability to Spring and Tabuchi (2021).

Tuning power was checked qualitatively by observing the number of times that students needed to repeat target words and phrases before successfully pronouncing the word well enough to receive a passing score from the NatTos tool. This was checked for students overall to observe the general effects, and representative examples of individual target words and phrases are also presented. We assume that if students can change their pronunciation from the first attempt to a final successful attempt, then it is sufficient evidence that the training and feedback from the tool had some tuning power. Furthermore, representative examples of student mistakes and changes amongst their attempts are provided to show how students adjusted their pronunciation in accordance with the feedback from the tool.

Results

Table 3 shows the results of the comparison of the pre- and post-test scores from this study. The average number of overall errors decreased significantly, indicating a medium effect size. Vowel sound errors, errors due to epenthesis, and the number of other errors also showed significant decreases, but with different effect sizes (small, medium, and large, respectively). However, there was not a significant decrease in any of the specific individual consonant sound errors observed in this study.

Table 3

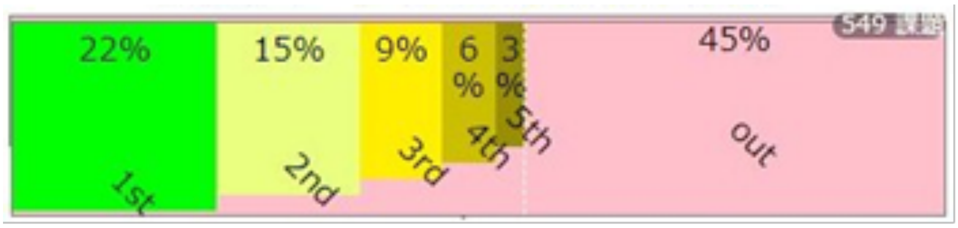
Comparison of the Average (sd) Number of Errors in Pre- and Post-tests

Error Type	Pre-test	Post-test	Statistical Comparison
Total Number	16.6 (7.7)	10.9 (5.5)	t = -2.6, p = .02, d = .85*
R/L Errors	3.1 (2.1)	3.6 (2.0)	t = -.39, p = 0.7, d = .13, n.s.
V/B Errors	0.2 (0.4)	0.2 (0.5)	t < .01, p = 1, d = 0, n.s.
TH/S/Z Errors	0.8 (1.0)	0.5 (0.6)	t = -1.24, p = .23, d = .39, n.s.
Vowel Sound Errors	4.4 (2.8)	2.7 (2.3)	t = -2.53, p = .02, d = .64*
Epenthesis Errors	0.9 (1.3)	0.2 (0.4)	t = -2.17, p = .04, d = .72*
Other Errors	7.6 (3.2)	4.7 (2.0)	t = -2.84, p = .01, d = 1.01*

Figure 2 shows the overall trend regarding which attempt students could successfully pronounce the target word or phrase. Generally, students could only pronounce a word or phrase correctly 22% of the time on the first try, but after repeated attempts, they were clearly able to adjust their pronunciation in many instances, as they could receive a passing score by at least the fifth attempt at least 55% of the time.

Figure 2

Overall Average Rate of Success During First through Fifth Attempts

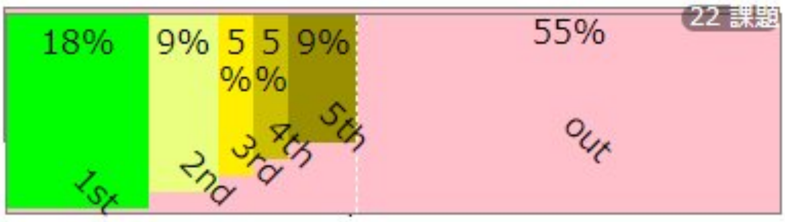


Note: "out" indicates non-success

A representative example of a single participant's attempts over a given set of target words in phrases from class 13 is shown in Figure 3. This participant (participant number 15, henceforth P15) was only successful on their first attempt 18% of the time, but successful before the fifth attempt 45% of the time.

Figure 3

P15's Success Rate for Class 13 Practice Drills



P15 attempted each set (12 target words and phrases) each, and then retried the target phrases that they missed ten times, resulting in a total of 22 attempts, which are all included in Figure 3. Table 3 shows the target words and phrases that P15 was attempting, which ones they were successful at more immediately, and which ones were more difficult. The data shows that P15 could successfully pronounce 33% of the target words and phrases on the first attempt and 67% within the first five attempts. The remaining four target words and phrases were quite difficult and attributed to a large number of "out" results in Figure 3. P15 was able to pronounce half of these difficult target words and phrases finally successfully after several attempts but was unable to pronounce "look for the railroad tracks" and "there are a great variety of animals" successfully.

Table 4

Number of Attempts P15 Made for Class 13 Target Words and Phrases Before Success

		Successful Attempt and Number of Repeats			
#	Target Word or Phrase	1	2	3	4
1	adorned with	F	F	5th	
2	hit hard times	F	F	F	2nd
3	Breathtaking	1st			
4	Goldfish	1st			
5	they live in water	3rd			
6	caught in a trap	2nd			
7	look for the railroad tracks	F	F	F	
8	hardly noticeable	1st			

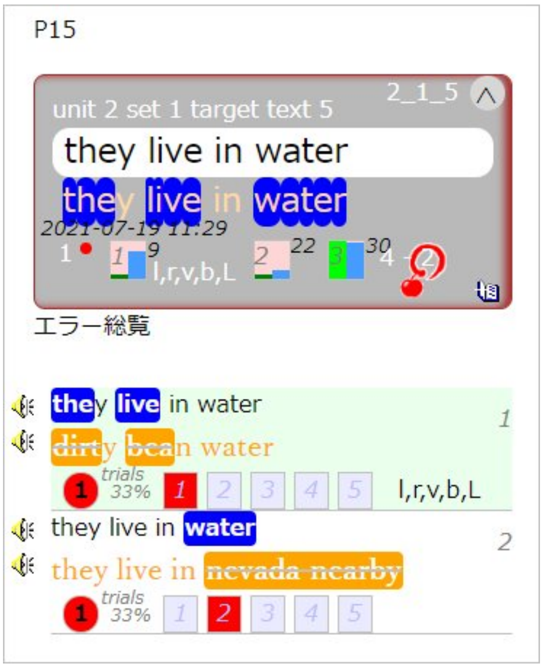
9	One of the most famous places is Central Park.	1st			
10	The land was bought a long time ago.	5th			
11	There are a great variety of animals.	F	F	F	F
12	If you were to walk from one end to the other	4th			

Note: "F" indicates all five attempts were failures. P15 did not repeat target words and phrases that they successfully pronounced within the first five attempts.

The results of Table 4 indicate that the feedback and training given by the NatTos tool were able to guide P15 to an intelligible pronunciation of 50% of the target words and phrases, i.e., those that they did not successfully pronounce the first time, but could eventually pronounce after practice. To further observe this process, a representative example of P15's changes in pronunciation over the course of their attempts, i.e., their three attempts at "they live in water," is presented in Figure 4.

Figure 4

Three Attempts to Pronounce "they live in water" by P15



P15

unit 2 set 1 target text 5

they live in water

2021-07-19 11:29

エラー総覧

they live in water

~~dirt~~y ~~bea~~n water

they live in water

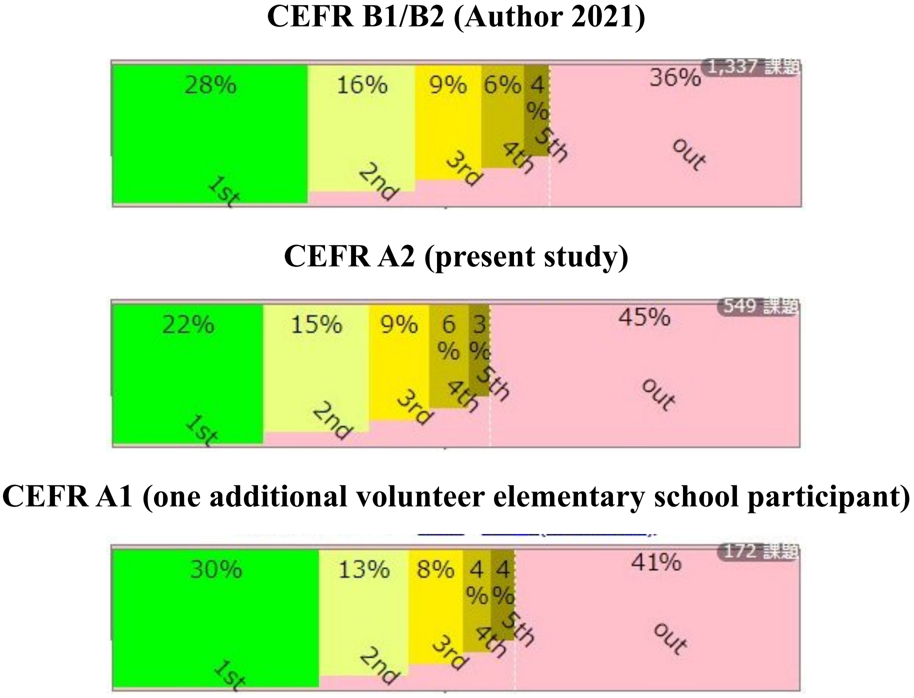
they live in ~~nevada nearby~~

Figure 4 shows that P15's first attempt resulted in the ASR recording "dirty bean water," suggesting that they likely pronounced *water* sufficiently well, but that they had difficulty with the words *they* and *live*. The errors that caused the mispronunciation are likely a combination of (1) pronouncing *ð* as *d* resulting in a word beginning with a *d* instead of with a *th*, i.e., *dirty* instead of *they*, and (2) epenthesis resulting in a two-syllable word in place of a one-syllable word, i.e., *dirty* instead of *they*, and (3) a mispronunciation of *ɪ̞* as *i*, resulting in *bean* instead of *live in*. P15 clearly focused on improving their pronunciation of these words, as their second attempt resulted in an acceptable pronunciation of the words, *they live in*. Furthermore, P15 spoke at a speed of 2.3 words per second during their first attempt, but at a speed of 0.7 words per second during their second attempt, suggesting they were focusing much more carefully on pronunciation at the expense of fluency. However, while focusing on these words, it seems that P15 still had trouble with epenthesis, causing the two-syllable word *water* to be recognized as a multiple word expression *Nevada* nearby. Since Nevada nearby contains the necessary sounds that are required for water and yet difficult for Japanese speakers, i.e., *ð*, as is generally pronounced in American English and *r*. On the third attempt, P15 was able to correct their pronunciation of all of the words and sounds in question, resulting in an acceptable pronunciation of the target phrase. Furthermore, their third attempt was conducted at a speed of 3.0 words per second, suggesting that they had remedied their pronunciation enough that they did not need to focus on it as much and could once again return to a more normal rate of speaking. This progression suggests that P15 changed their focus based on the feedback they received and was able to successfully alter their pronunciation through the training and feedback provided by the tool. Furthermore, the results suggest that epenthesis may be particularly difficult for P15.

Finally, we compared the average rate of success during the first through fifth attempts of drills for this current study's data set and that of Spring and Tabuchi (2021), as shown in Figure 5. We noticed a similar amount of gradual improvement over repeated attempts, despite the difference in the L2 proficiency levels of the participants in this study and Spring and Tabuchi (2021); CEFR A2 and CEFR B1/B2, respectively. Since L2 proficiency level can impact pronunciation acquisition (e.g., Chau et al., 2022; Glonka et al., 2014; Spring & Tabuchi, 2021), we also recruited one volunteer elementary school participant who participated in drills and was at CEFR A1 level. A comparison of these three shows that the trend holds at all levels.

Figure 5

Overall Average Rate of Success During First through Fifth Attempts Across Ability Levels



CEFR B1/B2 (Author 2021)

CEFR A2 (present study)

CEFR A1 (one additional volunteer elementary school participant)

Discussion

Taken summatively and in consideration of other studies, the results of this study offer some important insights. First, we found a similar amount of overall improvement in pronunciation ability as Spring and Tabuchi (2021), albeit slightly less than the "mid-level" group reported in Author (2021). This result in itself is not particularly surprising, as a number of studies that have considered computer-assisted pronunciation also report that using it helped improve students' pronunciation (e.g., Bozorgian & Shamsi, 2020; Guskaroska, 2019; 2020; Wang & Crosthwaite, 2021). What is somewhat more interesting are the differences in which areas were most improved.

While Spring and Tabuchi (2021) reported that learners improved the most on consonant sounds and that these improvements had the greatest impact on improving intelligibility, this study found that participants improved at the pronunciation of vowel sounds, including lowering the amount of epenthesis than any one specific consonant sound. The reasons for this seeming difference in results could be due to a number of reasons. First, Spring and Tabuchi (2021) analyzed their results based on wide categories, such as “consonant sounds” and “vowel sounds,” but this study divided consonant sounds into specific error types and divided “vowel sounds,” into differences in individual vowel sounds and problems with epenthesis. This allowed us to check more specifically which pronunciation errors were more prevalent and which improved. For example, we found that, specifically, the *r* and *l* pronunciation error, long reported as pervasive in L1 Japanese EFL speakers (e.g., Goto, 1971), did not show improvement, but errors such as the difference between *θ* and *s* and between *ð* and *z* did show some improvement. Furthermore, we were able to find that both vowel sound and epenthesis errors improved through the lessons and training. However, it should be noted that the learners in this study exhibited very few consonant errors outside of *r* and *l* during their pre-tests, and so the results might also simply be because university-level Japanese EFL learners have more difficulty with vowel sounds and *r* and *l* sounds as opposed to other consonant errors commonly associated with them, such as *v* and *b*, *θ* and *s*, and *ð* and *z*. Furthermore, it could be that though learners are still making some mistakes with many consonant sounds, these areas of pronunciation simply cause fewer problems with intelligibility than mispronouncing vowels or *r* and *l* sounds.

Another reason for differences in improvement could be because Spring and Tabuchi (2021) reported on intensive pronunciation instruction over five sessions that included ASR training, whereas this study reports on less intensive pronunciation instruction that lasted over a longer period of time and included more ASR training sessions. Since both studies showed similar amounts of improvement, it does not seem that the length of the sessions and training had an impact on improvement in overall intelligibility, but they might have had an impact on which particular pronunciation points are more likely to improve, i.e., it could be that intensive instruction and training might be more impactful on consonant pronunciation whereas less intensive but prolonged training might be more impactful on vowel pronunciation. However, more study is required to know if any of these factors contributed to differences in the results or if any differences are simply due to statistical variance.

Finally, differences in improvement in this study and Spring and Tabuchi (2021) might be the beginning levels of the students. Spring and Tabuchi (2021) found differences in improvement depending on beginning pronunciation levels, but the participants in that study were mostly CEFR B1 and B2 levels, whereas the participants in this study could best be described as CEFR A2. However, since Figure 5 shows a similar trend of gradual improvement on repeated attempts regardless of proficiency level, we suggest that ability level likely does not affect overall improvement through the pronunciation instruction and ASR-based training, but rather has an impact on which particular pronunciation points learners are struggling with or learning to overcome.

Another finding of this study was that ASR-based pronunciation training can clearly play an important role in the learning of L2 pronunciation. Though pronunciation instruction is also likely helpful to students, Figure 2 and Figure 3 clearly show that guided practice also plays a role in improvement, as students were often unable to pronounce target words and phrases correctly on their first attempt, despite having just received instruction and teacher-guided practice. However, Figure 2 also suggests that some troubles persisted despite instruction, feedback, and practice. This could be due to several reasons. For one, some target words or phrases might simply be difficult for ASR to guess if the frequency is too low and not enough context is provided. Therefore, instructors must carefully select target words and phrases and work to embed them in just the right amount of context so that the correctness of the pronunciation will be correlated with the accuracy score of the tool (see Ashwell & Elam, 2017; Inceoglu et al., 2020; Spring, 2020). Another reason that participants may not have been able to produce a satisfactory pronunciation may have to do with the length of the target word or phrase. Longer words and phrases will increase cognitive load and lessen the chance that learners can function on specific pronunciation points with which they are having trouble. However, the representative results shown in Table 3 suggest that this is not necessarily the case, as P15 was able to successfully pronounce several longer target phrases after multiple attempts, whereas sometimes it was a shorter phrase that caused them difficulty. Finally, it could also be the case that certain learners have more difficulty with particular sounds. While this is well documented based on L1 (e.g., Goto, 1971; Mochizuki, 1981; Nakata & Shockey, 2011), such difficulties could also be due to learner level and individual differences in learning styles or personal pronunciation, such as regional dialect.

The results of this study help elucidate the process by which students improve their pronunciation during ASR-based training. Specifically, the qualitative analysis shows a representative pattern by which a student makes a pronunciation error, then adjusts their pronunciation based on the feedback given by the tool. As represented by the example of P15, learners tended to focus heavily on their pronunciation after failing an initial time, causing them to sacrifice fluency for accuracy, as predicted by the trade-off hypothesis (Skehan & Foster, 1997). However, we found that learners then tended to increase their speaking speed on subsequent attempts, even as they pronounced the target words more clearly, which suggests that such training allows for improvements in multiple areas of spoken proficiency, i.e., both accuracy and fluency, as suggested by works such as Vercelloti (2017). It is unclear if this is simply a universal pattern of acquisition or if this quick return to initial fluency but with better pronunciation is due to the intervention of the NatTos tool, but this is worthy of further study as well.

Conclusion

In conclusion, this study was able to show that text-based ASR-based feedback and pronunciation training has a discernable impact on the intelligibility of students outside of the pronunciation instruction that is usually integrated into such studies. Furthermore, the results suggest that instruction and training can be either intensive or extensive and performed across a wide variety of ability levels with similar outcomes in terms of improved intelligibility. Finally, it suggests that L1 Japanese EFL learners have more difficulty improving some pronunciation points, i.e., *r* and *l*, than others, i.e., vowel sounds and epenthesis. However, this study also illuminated the need for a number of further studies. Specifically, the results suggest a need to evaluate which specific pronunciation points are more difficult to improve than others for specific groups of L1-L2 pairings. Furthermore, it seems possible that learners might only temporarily trade off fluency for pronunciation accuracy, but quickly be able to retain both, but this requires further study. Finally, as with many other learning studies, there seems to be a large degree of individual differences amongst learners that should be explored further, particularly in studies that examine the effect of peer-based feedback used in conjunction with ASR-based feedback (e.g., Dai & Wu, 2021; Wu et al., 2022). Future studies should also work to discover why some learners improve at certain pronunciation points more than others and what factors contribute to these differences. Some possibilities include the model text-to-speech voice that learners used when listening, L1 pronunciation styles including dialects, and the length of the practice words and phrases used in the practice sessions.

References

Ahn, T., & Lee, S. M. (2016). User experience of a mobile speaking application with automatic speech recognition for EFL learning. *British Journal of Educational Technology, 47*(4), 778–786. https://doi.org/10.1111/bjet.12354

Ashwell, T., & Elam, J. R. (2017). How accurately can the Google Web Speech API recognize and transcribe Japanese L2 English learners' oral production? *JALT CALL Journal, 13*(1), 59–76. https://doi.org/10.29140/jaltcall.v13n1.212

Bozorgian, H., & Shamsi, E. (2020). Computer-assisted pronunciation training on Iranian EFL learners' use of suprasegmental features: A case study. *Computer-Assisted Language Learning Electronic Journal, 21*(2), 93–113. http://callej.org/journal/21-2/Bozorgian-Shamsi2020.pdf

Chau, T., Huensch, A., Hoang, Y. K., & Chau, H. T. (2022). The effects of L2 pronunciation instruction on EFL learners' intelligibility and fluency in spontaneous speech. *TESL-EJ, 25*(4). https://tesl-ej.org/pdf/ej100/a7.pdf

Chen, H. N. (2011). Developing and evaluating an oral skills training site supported by automatic speech recognition technology. *ReCALL Journal, 23*(1), 59–78. https://doi.org/10.1017/s0958344010000285

Dai, Y., & Wu, Z. (2021). Mobile-assisted pronunciation learning with feedback from peers and/or automatic speech recognition: A mixed-methods study. *Computer Assisted Language Learning*, https://doi.org/10.1080/09588221.2021.1952272

Evers, K., & Chen, S. (2020). Effects of an automatic speech recognition system with peer feedback on pronunciation instruction for adults. *Computer Assisted Language Learning*, 1–21. https://doi.org/10.1080/09588221.2020.1839504

Golonka, E. M., Bowles, A. R., Frank, V. M., Richardson, D. L., & Freynik, S. (2014). Technologies for foreign language learning: A review of technology types and their effectiveness. *Computer Assisted Language Learning, 27*(1), 70–105. https://doi.org/10.1080/09588221.2012.700315

Goodine, M. (2019). TOEFL speaking 2019 sample TOEFL speaking questions and answers. https://www.toeflresources.com/speaking-section/toefl-speaking-samples/

Goto, H. (1971). Auditory perception by normal Japanese adults of the sounds "L" and "R". *Neuropsychologia, 9*, 317–323. https://doi.org/10.1016/0028-3932(71)90027-3

Guskaroska, A. (2019). ASR as a tool for providing feedback for vowel pronunciation practice. [Master's thesis, Iowa State University]. Iowa State University Digital Repository. https://dr.lib.iastate.edu/server/api/core/bitstreams/beac9ba0-733d-4f6b-ba44-9385e479fc32/content

Guskaroska, A. (2020). ASR-dictation on smartphones for vowel pronunciation practice. *Journal of Contemporary Philology, 3*(2), 45–61. https://doi.org/10.37834/JCP2020045g

Hsu, L. (2015). An empirical examination of EFL learner's perceptual learning styles and acceptance of ASR-based computer assisted pronunciation training. *Computer Assisted Language Learning, 29*(5), 881–900. https://doi.org/10.1080/09588221.2015.1069747

Hu, G., & Su, J. (2015). The effect of native/non-native information on non-native listeners' comprehension. *Language Awareness, 24*, 273–281. https://doi.org/10.1080/09588221.2015.1069747

Inceoglu, S., Lim, H., & Chen, W. (2020). ASR for EFL pronunciation practice: Segmental development and learners' beliefs. *The Journal of Asia TEFL, 17*(3), 824–840. http://dx.doi.org/10.18823/asiatefl.2020.17.3.5.824

Kim, D., & Kim, D. J. (2012). Effect of screen size on multimedia vocabulary learning. *British Journal of Educational Technology, 43*(1), 62–70. https://doi.org/10.1111/j.1467-8535.2010.01145.x

Levis, J. (2018). *Intelligibility, oral communication, and the teaching of pronunciation.* Cambridge University Press.

Lindemann, S., & Subtirelu, N. (2013). Reliably biased: The role of listener expectation in the perception of second language speech. *Language Learning, 63*(3), 567–594. https://doi.org/10.1111/lang.12014

McCrocklin, S. (2019a). ASR-based dictation practice for second language pronunciation improvement. *Journal of Second Language Pronunciation, 5*(1), 98–118. https://doi.org/10.1075/jslp.16034.mcc

McCrocklin, S. (2019b). Dictation programs for second language pronunciation learning: Perceptions of the transcript, strategy use and improvement. *Konin Language Studies, 7*(2), 137–157. https://doi.org/10.30438/ksj.2019.7.2.2

McCrocklin, S., & Edalatishams, I. (2020). Revisiting popular speech recognition software for ESL speech. *TESOL Quarterly, 54*(4), 1086–1097. https://doi.org/10.1002/tesq.3006

Mehrpour, S., Shoushtari, S. A., & Shirazi, P. H. N. (2017). Computer-Assisted Pronunciation Training: The Effect of Integrating Accent Reduction Software on Iranian EFL Learners' Pronunciation. *Computer-Assisted Language Learning Electronic Journal, 17*(1), 97–112. http://callej.org/journal/17-1/Mehrpour-Shoushtari-Shirazi2016.pdf

Mochizuki, M. (1981). The identification of /r/ and /l/ in natural and synthesized speech. *Journal of Phonetics, 9*, 283–303. https://doi.org/10.1016/S0095-4470(19)30972-6

Mroz, A. (2018). Seeing how people hear you: French learners experiencing intelligibility through automatic speech recognition. *Foreign Language Annals, 51*(3), 617–637. https://doi.org/10.1111/flan.12348

Munro, M. J. (2010). Intelligibility: Buzzword or buzzworthy? In J. Levis & K. LeVelle (Eds.), *Proceedings of the 2nd pronunciation in second language learning and teaching conference*, (pp. 7–16). Iowa State University. https://www.academia.edu/12031228/Proccedings_of_the_2nd_Pronunciation_in_Second_Language_Learning _and_Teaching_Conference

Nakata, H., & Shockey, L. (2011). The effect of singing on improving syllabic pronunciation – Vowel epenthesis in Japanese. *Proceedings of the 17th International Congress of Phonetic Sciences* (ICPhS), 1442-1445. https://www.internationalphoneticassociation.org/icphsproceedings/ICPhS2011/OnlineProceedings/RegularSession/Nakata/Nakata.pdf

Neri, A., Mich, O., Gerosa, M., & Giuliani, D. (2008). The effectiveness of computer-assisted pronunciation training for foreign language learning by children. *Computer Assisted Language Learning, 21*(5), 393–408. https://doi.org/10.1080/09588220802447651

Otake, T., Hatano, G., Cutler, A., & Mehler, J. (1993). Mora or syllable? Speech segmentation in Japanese. *Journal of Memory and Language, 32*, 258–278. https://doi.org/10.1006/jmla.1993.1014.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878–912. https://doi.org/10.1111/lang.12079

Sidgi, L. F. S., & Shaari, A. J. (2017). The usefulness of automatic speech recognition (ASR) Eyespeak software in improving Iraqi EFL students' pronunciation. *Advances in Language and Literary Studies, 8*(1), 1–6. https://doi.org/10.7575/aiac.alls.v.8n.1p.221

Skehan, P., & Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research, 1*, 185–211. https:/doi.org/10.1177/136216889700100302

Spring, R. (2020). Using multimedia tools to objectively rate the pronunciation of L1 Japanese EFL learners. *ATEM Journal: Teaching English through Multimedia, 25*, 113–124.

Spring, R., & Tabuchi, R. (2021). Assessing the practicality of using an automatic speech recognition tool to teach English pronunciation online. *STEM Journal, 22*(2), 93–104. https://doi.org/10.16875/stem.2021.22.2.93

Vercellotti, M.L. (2017). The development of complexity, accuracy, and fluency in second language performance: A longitudinal study. *Applied Linguistics, 38*(1), 90–111. https:/doi.org/10.1093/applin/amv002

Wallace, L. (2016). Using Google Web Speech as a springboard for identifying personal pronunciation problems. In J. Levis, H. Lee, I. Lucic, E. Simpson, & S. Vo (Eds.), *Proceedings of the 7th annual pronunciation in second language learning and teaching conference* (pp. 180–186). Iowa State University.

Wang, Y. S., & Young, S. S.-C. (2012). Exploring young and adult learners' perceptions of the corrective feedback in the ASR-based CALL system. *British Journal of Educational Technology, 43*(3), 77–80. https://doi.org/10.1111/j.1467-8535.2011.01275.x

Wang, Y. S., & Young, S. S.-C. (2015). Effectiveness of feedback for enhancing English pronunciation in an ASR-based CALL system. *Journal of Computer Assisted Learning, 31*(6), 493–504. https://doi.org/10.1111/jcal.12079

Wang, H., & Crosthwaite, P. (2021). The affordances of WeChat voice messaging for Chinese EFL learners during private tutoring. *Computer Assisted Language Learning Electronic Journal, 22*(1), 230–253. http://callej.org/journal/22-1/Wang-Crosthwaite2021.pdf

Wu, X., Liu, X., & Chen, L. (2022). Reducing EFL learners' error of sound deletion with ASR-based peer feedback. In W. Jia, Y. Tang, R. S. T. Lee, M. Herzog, H. Zhang, T. Hao, & T. Wang (Eds.) *Emerging technologies for education. SETE 2021* (pp. 178–189). Springer. https://doi.org/10.1007/978-3-030-92836-0_16

Xiao, W., & Park, M. (2021). Using automatic speech recognition to facilitate English pronunciation assessment and learning in an EFL context: Pronunciation error diagnosis and pedagogical implications. *International Journal of Computer-Assisted Language Learning and Teaching, 11*(3), 74–91. https://doi.org/10.4018/IJCALLT.2021070105