*Journal of English Teaching through Movies and Media*
Vol. 22, No. 2, (May 31, 2021), pp. 93-104
https://doi.org/10.16875/stem.2021.22.2.93

ISSN 2765-7078
https://www.ejournal-stem.org

# Assessing the Practicality of Using an Automatic Speech Recognition Tool to Teach English Pronunciation Online

**Spring, Ryan[1]; Tabuchi, Ryuji[2]**

## Abstract

This study aims to determine how well an automatic speech recognition (ASR) tool could be used in an online EFL course to help L1 Japanese students improve their pronunciation. Previous studies have suggested that ASR tools can be helpful in this regard, but few have dealt with an entirely online course or attempted to use the data to determine which lessons were the most impactful. Therefore, we used a mixture of pre- and posttest data and survey results to determine how students would receive the ASR tool, whether or not they would improve, and what lessons were the most useful to them. The results suggest that students were generally positive towards the ASR-assisted practice and that they quite clearly improved their intelligibility, especially students who began with lower ability. Specifically, the tool was found to be most useful for students who had a lower than 95% accuracy on their pretests. Furthermore, students claimed the tool was most helpful for practicing consonant and vowel sounds, but a statistical model was unable to pinpoint which lessons were most helpful to their overall improvement. Future work should aim to discover the best drills to use and which pronunciation points to teach.

**Keywords:** automatic speech recognition (ASR) tools, pronunciation, EFL, intelligibility, online teaching
**Applicable levels:** elementary, secondary, tertiary

* This study was approved by the Research Ethics Review Boards (IRB) of the Tohoku University Institute for Excellence in Higher Education (No. k0213).

[1] Corresponding author, Associate Professor, Institute for Excellence in Higher Education, Tohoku University, Aoba-ku, Kawauchi 41, Miyagi Prefecture, Sendai City, Japan (E-mail: spring.ryan.edward.c4@tohoku.ac.jp)
[2] Co-author, Owner, Mint Phonetics Education Institution, Hagiwara-machi 950-31, Gunma Prefecture, Takasaki City, Japan (E-mail: tabuchiryuji@nifty.ne.jp)

Received: March 18, 2021
Revised: May 11, 2021
Accepted: May 23, 2021

## I. INTRODUCTION

One aspect of acquiring any second language is learning the pronunciation of the target language. While perfect pronunciation may not be necessary for learners, they must be able to produce speech that is at the very least understood by others, or else it is impossible for communication to occur. However, despite how important it is, pronunciation can be very difficult in comparison to other aspects of language learning because of differences in the first (L1) and second language (L2) of the learners, and because it is difficult for students and teachers to know how good one's pronunciation has to be in order for them to be understood.

The difficulty in learning pronunciation depends on a wide range of factors including how similar the phonetic structure of the learners' L1 and L2 are. In the case of L1 Japanese EFL learners, it is especially difficult to master the L2 pronunciation because the two languages are so phonologically different. For example, English is generally thought to contain around 24 different consonant phonemes and up to 20 distinct vowel phonemes (though not all are used in all dialects), whereas Japanese contains just 15 consonant phonemes and 5 vowels (Rogers, 2000; Tsutomu, 2000). Furthermore, English is said to be a syllabic language whereas Japanese is mora-based, which causes differences in rhythm, stress, and timing (Otake, Hatano, Cutle, & Mehler, 1993). With all of these differences, it can be difficult to properly imitate the pronunciation of native speakers, but it also leaves Japanese EFL students wondering which aspects of pronunciation are the most important in order for them to be understood.

Another problem when teaching and learning L2 pronunciation is determining how good one's pronunciation must be in order to be understood. While a number of studies and teaching techniques tend to focus on subjective judgements of students' pronunciation, either by native-speaking judges or the teachers themselves, a number of studies have pointed out, such judgements are highly influenced by the listener's perception of the speaker and can thus be flawed (e.g., Hu & Su, 2015; Lindemann & Subtirelu, 2013). Furthermore, subjective judgements are often focused more on comprehensibility and interpretability, but Munro (2010) suggests that focus should instead be placed on intelligibility, i.e., how well an L2 speaker's utterances can be understood. Here, advances in automatic speech recognition (ASR) software can be useful, as the software uses context clues, somewhat similar to how a human might, and attempts to recreate a speaker's speech in text format. Therefore, using such tools can help teachers and researchers to better monitor the intelligibility of students' pronunciation as it is objective and gives instructors and researchers a clue as to how well the speaker would be understood by a native speaker (e.g., Ashwell & Elam, 2017; Spring, 2020).

Learning proper English pronunciation is also made difficult for Japanese EFL learners by the fact that students are often not given enough time for speaking practice, and generally have a lack of motivation and willingness to communicate (WTC) in the classroom (Yashima, 2002). Additionally, the lack of practical speaking practice has only been further exacerbated by the COVID-19 pandemic, which forced many universities and schools in Japan to suddenly shift to attempting to teach EFL classes online. Even if teachers utilize real-time teaching tools such as synchronous video mediated communication, it is difficult for teachers to listen to the pronunciation of several students and give individualized feedback. Here again, ASR technology can potentially help, as it allows for students to speak into a number of increasingly common devices (e.g., computers and smartphones) and instantaneously have the intelligibility of their pronunciation evaluated.

However, despite the use of some ASR integrated pronunciation tools (e.g., English Central, Duolingo, Eyespeak, etc.), it is still unclear how well ASR works to actually improve specific aspects of learner intelligibility, how instructor feedback can be helpful when using these tools, and how these can be integrated into online EFL teaching in the present day. Therefore, this study seeks to examine if ASR integrated practice tools are applicable in a completely online EFL learning environment, if certain aspects of Japanese EFL learners' intelligibility are aided at all by such tools, and if so, in what ways they are positively impacted.

## II. LITERATURE REVIEW

A number of studies have suggested that ASR-based computer mediated language learning tools can have a positive impact on L2 learners' acquisition of pronunciation. For example, studies such as Doremalen, Boves, Colpaert,

Cucchiarini, and Strik (2015) and Ahn and Lee (2016) found that both teachers and students were generally positive towards their usage. Furthermore, Hsu (2015) found that there was no relationship between learning style and the perceived usefulness of ASR-based computer assisted pronunciation training, indicating that it can be used positively with a wide variety of learners. Finally, Golonka, Bowles, Frank, Richardson, and Freynik (2014) examined a number of similar studies finding that overall ASR-based pronunciation training is reasonably beneficial to L2 learners.

However, despite the seeming benefits of ASR-based pronunciation training, there is still room for improvement. For example, though a number of studies have found that users are generally positive towards using ASR tools in EFL contexts, most show a satisfaction rate over 50%, but far from 100% (e.g., Guskaroska, 2019; Hsu, 2015; Sidgi & Shaari, 2017). One potential reason for this is the fact that most feedback from ASR comes only from the computer in an automatized way, which is likely due to the fact that learners generally use these tools outside of class or for individualized practice (Ahn & Lee, 2016). This could lead to students losing motivation as it lacks a human touch, and as pointed out by Doremalen et al. (2015) it is important for instructors to make sure that students remain motivated to use ASR-based tools. Furthermore, students may question whether or not the ASR is a good judge of pronunciation if the sentences, words, and phrases used for practice are not carefully chosen (Ashwell & Elan, 2017), as most ASR software uses n-grams or contextual clues to help it guess the correct word and form, and a lack of such information could lead it to mark homonyms as incorrect or bad pronunciation.

Although ASR training studies have been found to be useful in improving overall pronunciation in general (e.g., Golonka et al., 2014; Mroz, 2018), fewer studies have looked at how it can help specific pronunciation points. One exception is Guskaroska (2019) who examined the use of ASR to promote the acquisition of EFL vowel sounds, but found that while learners' overall pronunciation improved via the ASR assisted training, their ability to pronounce the targeted vowels did not. One reason that it is difficult to determine exactly which aspects of L2 pronunciation ASR-based tools can help learners to acquire is that learners with different L1s and L2s have various problems based on the phonological variations in their native and target language. In the case of L1 Japanese EFL learners, there are a number of challenges. The first is the problem of individual consonant and vowel sounds. Because English has far more phonemes than Japanese, L1 Japanese EFL learners often tend to confuse and conflate the following sounds: *r* and *l*, *v* and *b*, *θ* and *s*, *ð* and *z*, *æ* and *a*, *ɪ* and *i*, *ʌ* and *ɒ* and *a*, *ä* and *o* (e.g., Dolman & Spring, 2014; Goto, 1971; Mochizuki, 1981; Saito, 2011). Furthermore, English is a stress-based language, whereas Japanese is not, which means that it is difficult for L1 Japanese learners to recognize the difference in the pronunciation of the verb (rɪˈkɔːrd) and noun (ˊrekərd) form of words such as *record*. Finally, English makes great use of linking (e.g., *black coffee* is generally pronounced *blækɔːfi* as though it were one word) and reduction (e.g., *going to* is often pronounced *gənə*) in pronunciation, which can make it difficult for L1 Japanese EFL learners to acquire proper rhythm and intonation (Saito & Saito, 2016).

Furthermore, though much research on the use of ASR tools for pronunciation training generally checks improvements in pronunciation, the effects on fluency are unknown, as this is often not examined in such studies. It could be that being overly focused on pronunciation may improve intelligibility at the cost of speaking speed, as suggested by Skehan's Limited Capacity Hypothesis (1998), which suggests that focusing on one component of speech (complexity, accuracy, fluency) may result in lower performance on the other aspects. Training on individual sounds (e.g., consonants and vowel phonemes) surely does not help a learner to speak more quickly, and in fact, may cause a dip in fluency due to trade-off effects. However, if learners acquire a better mastery of reduction, linking, and rhythm, this may help them to pronounce individual words and phrases more smoothly, which could result in better fluency. However, this is currently unknown as most ASR studies tend to focus on the improvement of individual phonemes (e.g., Guskaroska, 2019; Mroz, 2018).

Finally, there is a question as to how the ASR-based pronunciation tools can be implemented in completely online courses. The aforementioned studies were all conducted with students who were in face-to-face teaching situations. However, the COVID-19 pandemic caused many instructors to suddenly be forced to teach through distance learning. Since it is nearly impossible for teachers to listen to each students' pronunciation, utilizing ASR could be a good way for students to still receive practical, individualized feedback, but there are no studies that can fully predict how students will receive ASR-based training when it is conducted through an online teaching environment.

This study seeks to offer insight beyond past works by observing student's reaction to ASR pronunciation tools in an online EFL class, and by creating an ASR-based pronunciation training tool in which learners can speak into

smartphones or computers and receive feedback not only from the automatized response, but also from their instructor, who can monitor students' responses and continue to give them encouragement. It examines a class of L1 Japanese university students learning English as a foreign language who utilized the tool during five pronunciation lessons (consonant sounds, vowel sounds, stress, linking and reduction, and rhythm) and seeks to answer the following research questions:

1. Is an ASR-based pronunciation training tool applicable for completely online EFL courses?
2. Do L1 Japanese EFL students improve their intelligibility and fluency through the use of the ASR-based pronunciation tool with teacher intervention?
3. How did the ASR-based training of consonant sounds, vowel sounds, stress, linking and reduction, and rhythm each impact students' overall pronunciation and fluency?

## III. METHOD

### 1. Participants

98 L1 Japanese first-year students learning English as a foreign language at a university in Japan volunteered to participate in this study. They were between the ages of 18 and 19 and had six and a half years of formal English study prior to the class. The participants were taking one English speaking and listening class given by the same instructor and one English reading and writing class given by three different instructors. The only pronunciation training that they received during this time was from the instructor of their English speaking and listening class. The students were a mix of CEFR A2, B1, and B2 students, as indicated by their TOEFL ITP scores: 400-640, $M = 524$, $SD = 44.8$. They participated in pronunciation training as a part of their English speaking and listening class, but they were asked for consent to use their data as part of this study and participated in the survey voluntarily, in line with the ethics review board of the students' University.

### 2. ASR Tool

The ASR tool used in this study, NatTos, was created in HTML, uploaded to a homepage, and made available through the internet. Instructors, as users, can create their own drills (words or sentences for students to practice) and then upload them for students to use. Students click on each drill, and are then presented with the written word or sentence. The word or sentence is automatically spoken vis a TTS (text to speech) algorithm, after which the students click the red button. Upon clicking, the program accesses the user's microphone and the user is asked to pronounce the word or sentence. The program activates the local ASR of the computer, device, or web application, and transcribes what the user said. The original word or sentence is then compared to the ASR transcription and the user is given a score based on how closely the transcription matched the pre-programmed word or sentence. The program presents what the ASR guessed that the user said, and if only part of the word was pronounced incorrectly, that area is shown in red to indicate the sound or area that the user made a mistake with. For example, if an instructor creates the drill *I don't eat rice*, and the ASR transcribes *I don't eat lice* for a user, the incorrect sentence that the ASR transcribed (i.e., *I don't eat lice*) is shown, and the letter *l* in *lice* is highlighted in red so that the user is alerted to the fact that this is the area in which they had a pronunciation error. Images of the tool in use are shown below in Figure 1. The left-hand picture of Figure 1 shows the tool, ready for recording. The practice sentence is displayed and students can push the audio button to hear the sentence again or the red button to speak the sentence. The right-hand picture of Figure 1 shows the tool after a student has attempted to the drill. In this case, the speaker didn't pronounce *he* or *likes* correctly, making a mistake with the consonant in the former, and with several of the sounds in the latter. As can be seen from the middle image, NatTos can be used not only on computers, but also on tablet and smartphones as well.

If users do not correctly pronounce the word or sentence, they are prompted to try again, up to five times, after which time the program automatically moves to the next drill. The user data is simultaneously sent to the creator of the drill (instructor) who can check the most common errors and problems for each drill in real time.

**FIGURE 1**

**NatTos ASR Tool in Use**


## 3. Procedure

The participants took a pretest of pronunciation before receiving instruction, and an identical posttest one week after the last training session. The task in both the pre- and posttest consisted of reading a set script of 200 words aloud with the clearest pronunciation they could during the last few minutes of class time. The script was based on a sample answer to a simple agree/disagree style question[1] that is asked on the TOEFL iBT test (Goodine, 2019), but modified to include consonant sounds, vowel sounds, word stress, linking and reduction, and rhythm pronunciation that was surmised to be difficult for L1 Japanese EFL learners. Students were asked to record their responses and submit them by the end of class. A simple survey was also given before and after training which consisted of questions related to their perceptions of English pronunciation (see Appendix). The accuracy of their pronunciation was analyzed by using an ASR (YouTube's automatized subtitling system) and having a native speaker check the number of mistakes, after which the number of mistakenly pronounced words was divided by the total number of words, as this method has been shown to be highly correlated with subjective native speaker judgements (Guskaroska, 2019; Spring, 2020). Their fluency was also analyzed using Praat (Boersma & Weenink, 2019), which has shown to be reasonably accurate in measuring a variety of variables highly correlated with fluency including speech rate, articulation rate, and the total number of pauses (De Jong & Wempe, 2009).

One week after the pretest, students received instruction in their English speaking and listening class once per week on one of the five pronunciation points introduced in section 2 from the same teacher – one of the authors of this paper. At the beginning of each 90-minute session, students were asked to use the ASR tool to pronounce eight words and eight sentences that were representative of the pronunciation point that would be taught during the class. Based on the results of the class, the instructor modified the lesson to focus on the points that the students seemed to be having the most difficulty with. Throughout the lesson, students were asked to use the ASR tool to pronounce other words and sentences representative of the pronunciation point. The instructor monitored their successes and mistakes and gave advice based on the mistake they had made. For example, in one class students were asked to pronounce the word *right*, but many students' ASR practice returned the word *white* instead. This led the instructor to know that students were moving their lips forward when pronouncing the *r* sound, and could advise them not to do so. During the last 10 minutes of the session, students were asked to use the ASR tool to pronounce the same words and sentences that they had attempted at the beginning of class. The scores for their pronunciations at the beginning and end of class were judged by NatTos and the numbers were used for later analysis. The five lessons (consonant sounds, vowel sounds, stress, linking and reduction, and rhythm) were based off of Chapter B2.5 in the university's textbook, *Pathways to Academic English* (Takebayashi, 2020).

Scores for the pre- and posttest pronunciation accuracy (as a percentage of words pronounced correctly), speech rate (syllables per second), articulation ratio (syllables per second), and the number of pauses were compared separately via dependent *t*-tests with Cohen's *d* given for effect size when significant differences were found. Scores for students' pronunciation on the drills used at the beginning and end of class were compared using the same analysis.

[1] The question was: "State whether you agree or disagree with the following statement. Then explain your reasons using specific details in your argument. Teachers should assign daily homework to students."

A multiple regression analysis was used to observe how well improvement (delta value) from the beginning to end of each lesson correlated to improvement in measures of pronunciation and fluency (delta value of pronunciation accuracy from the pre- and posttests).

Finally, a post-treatment survey was given to students to determine how well they received the ASR tool and the lessons in general. Two Likert-style questions were asked: one regarding whether or not they enjoyed using the tool, and one regarding whether or not they felt the tool helped them to improve. One multiple-selection style question was included asking students to select what areas of pronunciation they felt the tool helped them to improve. Finally, one open-ended question was used that simply asked for any other opinions regarding the pronunciation lessons and the ASR tool. Open-ended questions were conceptually clustered and representative comments were given for clarity.

## IV. RESULTS

### 1. Results of Measures of Improvement

Table 1 shows a comparison of pre- and posttest scores taken before and after the beginning of pronunciation training. Though the overall scores are given, it was discovered that several students began with very advanced speaking ability, so the pre- and posttest scores of students with high and mid-level speaking ability are also provided. Speaking ability level was determined by looking at the pronunciation accuracy of the overall pre-test and dividing students into the high level category if they achieved 95% or greater pronunciation accuracy, and the mid level category if they achieved less than a 95% accuracy score. A multiple factor repeated measures ANOVA was used to determine that there was a difference between the amount of improvement seen in mid- and high-level learners; $F = 32.45$, $p < .001$. *T*-test statistics for the pre- and posttest scores of learners are given with a Bonferroni correction.

**TABLE 1**

**Average (SD) Scores for Pre- and Posttests of Speaking**

| | Measure | Pretest | Posttest | *t*-test Statistics | | |
|---|---|---|---|---|---|---|
| | | | | *t* | *p* | *d* |
| Overall ($N = 98$) | Accuracy | 94.9% (.042) | 96.7% (.0199) | 32.45 | .03* | .45 |
| | Speech Rate | 3.09 (.40) | 3.07 (.34) | .90 | .37 | |
| | # of Pauses | 27.6 (8.25) | 29.3 (7.59) | -2.66 | .009** | .54 |
| | Art. Rate | 4.14 (.38) | 4.17 (.35) | -1.10 | .28 | |
| Mid-level ($N = 42$) | Accuracy | 91.71% (.046) | 95.96% (.021) | 5.35 | .001** | 1.17 |
| | Speech Rate | 3.02 (.39) | 3.05 (.37) | 1.39 | .17 | |
| | # of Pauses | 28.9 (7.37) | 29.7 (8.41) | .32 | .75 | |
| | Art. Rate | 4.07 (.37) | 4.16 (.37) | 2.08 | .04* | .45 |
| High-level ($N = 56$) | Accuracy | 97.23% (.012) | 97.26% (.017) | -.03 | .97 | |
| | Speech Rate | 3.15 (.40) | 3.08 (.32) | -1.68 | .10 | |
| | # of Pauses | 26.6 (8.8) | 29.1 (7.0) | 2.69 | .009** | .51 |
| | Art. Rate | 4.18 (.39) | 4.18 (.34) | -.10 | .92 | |

*Note.* *statistical significance of $p = .05$ or less, **significant at $p < .01$
Cohen's *d* only calculated for statistically significant results

In looking at just the overall results presented in Table 1, it would seem that overall accuracy improved only slightly, but the number of pauses went up, as indicated by the fact that a statistically significant difference was found between the accuracy and number of pauses of participants overall, and non-significant results for speech rate and articulation. However, in looking at the results divided by level, it would seem that mid-level learners improved their accuracy quite significantly, as indicated by the difference of over 4% between pre- and posttests and the very large effect size ($d = 1.17$) found. Furthermore, their articulation rates improved slightly as indicated by a statistically significant difference in pre- and posttest scores with a moderate effect size ($d = .45$), and without much change in speech rate or the number of pauses between the pre- and posttest scores ($p = .17$ and $p = .75$, respectively). Meanwhile, high-level learners did not seem to improve, as indicated by a lack of significant differences in similar variables between pre- and posttests, but rather did develop a tendency to pause more, as shown by an average increase of 2.5 pauses

from the pre- to posttest ($p$ = .009, $d$ = .51).

Table 2 shows the overall improvement exhibited in the pre- and post-class practice exercises for each lesson for students that completed 100% of both the pre- and post-class practice exercises. Unfortunately, some students had difficulty with the NatTos program or came late to class and therefore, pre- and post-class data is not available for each student for each lesson. Therefore, $N$ values are given for each individual test.

**TABLE 2**

**Pre and Post-Class ASR Tool Average (*SD*) Scores**

| | Lesson | Pre-class Score | Post-class Score | *t*-test Statistics | | |
|---|---|---|---|---|---|---|
| | | | | *t* | *p* | *d* |
| Overall | 1 ($n$ = 31) | 5.64 (1.1) | 5.95 (1.0) | 2.39 | .02* | .61 |
| | 2 ($n$ = 46) | 5.93 (1.2) | 6.07 (1.1) | .1.28 | .21 | |
| | 3 ($n$ = 13) | 7.68 (1.2) | 7.98 (1.4) | 1.31 | .21 | |
| | 4 ($n$ = 41) | 8.22 (.96) | 8.26 (.87) | .37 | .71 | |
| | 5 ($n$ = 46) | 6.93 (1.4) | 7.05 (1.4) | .84 | .41 | |
| Mid-level | 1 ($n$ = 10) | 5.05 (.8) | 5.50 (1.1) | 1.89 | .09 | |
| | 2 ($n$ = 16) | 5.12 (1.0) | 5.53 (1.0) | 1.78 | .09 | |
| | 3 ($n$ = 4) | 6.80 (0.9) | 6.85 (1.5) | $n$ < 8, Untestable | | |
| | 4 ($n$ = 14) | 7.69 (1.0) | 7.78 (0.9) | .36 | .73 | |
| | 5 ($n$ = 15) | 6.41 (1.3) | 6.37 (1.4) | -.19 | .85 | |
| High-level | 1 ($n$ = 19) | 5.98 (1.1) | 6.22 (.8) | 1.36 | .19 | |
| | 2 ($n$ = 30) | 6.37 (1.1) | 6.39 (1.0) | .16 | .88 | |
| | 3 ($n$ = 9) | 8.07 (1.1) | 8.49 (1.0) | 1.4 | .20 | |
| | 4 ($n$ = 26) | 8.53 (0.8) | 8.56 (0.7) | .14 | .89 | |
| | 5 ($n$ = 29) | 7.20 (1.4) | 7.46 (1.19) | 1.42 | .17 | |

*Note.* *statistical significance of $p$ = .05 or less, **significant at $p$ < .01
Cohen's $d$ only calculated for statistically significant results

The data presented in Table 2 seems to show that students overall showed some improvement directly after the first lesson, as indicated by a statistically significant difference in pre-class and post-class scores ($p$ = .02, $d$ = .61), but not directly after any of the other lessons, as evidenced by a lack of other statistically significant results. Furthermore, it does not suggest that mid-level or high-level students improved more or less from the beginning to end of classes. The reason for this is not entirely clear, but it could be due in part to the fact that the same students were not always represented each time.

## 2. Results of Correlation Between Lessons and Improvement

Table 3 shows the results of multiple regression analyses of the delta scores of the experimental pre- and posttests versus the deltas of the pre- and post-class practice exercises for each lesson. However, it should be noted that because most students had trouble with the NatTos ASR tool at the beginning of the third lesson, there was not enough available data to include it in the regression analyses. The multiple regression analysis was run for all students, and then again for both mid- and high-level students separately. The models for the overall data was statistically significant ($r^2$ = .58, $p$ = .05), but not for either level; $r^2$ = .95, $p$ = .1 (mid-level) and $r^2$ = .51, $p$ = .61 (high-level). Since neither of the models was significant, only the overall data was reported.

**TABLE 3**

**Results of Multiple Regression Analysis of Overall Delta Scores ($N$ = 40)**

**(change in overall pronunciation versus change in pronunciation after each lesson)**

| Lesson | Accuracy | | Speech Rate | | # of Pauses | | Art. Rate | |
|---|---|---|---|---|---|---|---|---|
| | *β* | *p* | *β* | *p* | *β* | *p* | *β* | *p* |
| 1 | .233 | .30 | .391 | .21 | -.486 | .11 | .207 | .49 |
| 2 | .218 | .33 | .104 | .73 | .267 | .36 | .400 | .19 |
| 4 | .743 | .02* | -.039 | .92 | .056 | .87 | -.035 | .92 |
| 5 | -.616 | .03* | -.133 | .72 | -.127 | .72 | -.304 | .77 |

*Note.* *Statistical significance of $p$ = .05 or less

According to the results presented in Table 3, students who improved from the beginning to the end of class on lesson 4 tended to improve overall, as indicated by a statistically significant result ($p$ = .02) and a positive beta value (.743). Conversely, those who improved in lesson 5 tended to get worse scores on their overall posttests, as evidenced by a statistically significant result ($p$ = .03) and a negative beta value (-.616). However, it should be noted that less than half of all participants submitted enough data to be considered in the model represented by Table 3, which has a large effect on the results. No other statistically significant results were found from this analysis, and therefore, further conclusions can not be drawn from this test.

### 3. Results of the Survey

In general, the results of the survey (respondents; $N$ = 76) seem to suggest that students were generally positive towards the NatTos ASR pronunciation tool. First, the students responded to the Likert-style question asking if they enjoyed using NatTos (1 = *I didn't enjoy it at all*, 5 = *I enjoyed it very much*) mostly positively, with an average score of 3.7 (median = 4). Secondly, the students also responded positively to the Likert-style question asking if they felt their pronunciation improved through using NatTos (1 = *I didn't improve at all*, 5 = *I improved a lot*), responding with an average score of 4.0 (median = 4). Furthermore, though only 15 students provided a free-response answer, six of them specifically mentioned that they felt it was a good way to practice pronunciation. Representative comments include, "[I didn't know that pronunciation tools like NatTos existed, but by doing it, I felt that my pronunciation improved, and it was fun]," and "[NatTos really taught me where my own pronunciation was wrong, which made it easy to understand where I was messing up]."

Regarding what areas they felt that NatTos helped them the most, students seemed to indicate that is was most useful to them for individual sounds. In response to the multiple-selection question (students were allowed to select as many answers as they wanted) asking which of the lessons NatTos helped them with the most, 61 (78%) of respondents said selected "vowel sounds," followed most closely by 59 students (75%) who selected "consonant sounds." In comparison, only 25 students (32%) selected "stress" and "rhythm," while only 2 (3%) selected "other" sounds. However, this could simply be due to the difficulty of these lessons, as evidenced by one student who simply wrote "[Rhythm was really difficult]" in the free-response section of the survey.

Finally, the survey also revealed some points that should be addressed in the future. Specifically, in the open response question, five students responded with doubts about NatTos' ability to correctly gauge their pronunciation. Representative responses included "I doubt if [NatTos] can surely record my voice because it can't recognize my voice even I speak loudly with little noise like air conditioner," and "The NatTos ASR accuracy wasn't very good. It often did not register the word "but" when I said it and it often registered "he" as "she" when speaking, which was really tough." As mentioned earlier in the paper, there were some problems with some of the practice sentences and drills selected for use in this study. Therefore, it will be critical to identify sentences and words that are best suited not only for pronunciation practice, but specifically for use with an ASR-based training tool. However, there may also be some problems with the actual pronunciation of students as well. Specifically, the student who mentioned that the word *he* was registered as *she* consulted the teacher towards the end of the class, but when speaking with the teacher, he confirmed that when the student said *he* it sounded like *she* to him, so perhaps the problem was not with the ASR tool, but with the students' pronunciation and the lesson materials. Since the difference in *hi:* and *ʃi:* is not well represented as a prevalent problem for L1 Japanese EFL learners in the literature, we did not consider it for inclusion in the lessons. However, if several students had similar problems, it might be worth teaching specifically in the future.

## V. DISCUSSION

Summatively, the results of this study seem to show that the lessons and practice conducted using the NatTos ASR tool were applicable to a completely online EFL class, and helpful in improving the intelligibility of students whose intelligibility began under 95% (as measured by how accurately ASR was able to transcribe a given text read by the participants). This was demonstrated in part by the results of the survey, which showed that students were grateful to

have such a tool during the online class and in general felt that it was fun and helpful. Student improvement was evident through the pre- and posttest scores of students with lower beginning intelligibility levels, a gain of nearly 5% in intelligibility, as measured by how accurately an ASR could predict participants' spoken words. According to Spring (2020), using the methods described in this paper to judge intelligibility, an increase of between 5% and 7% in the ability of an ASR to correctly predict a second language learners' spoken words generally marks a noticeable difference as judged by native speakers, by about one point on a ten-point scale. Therefore, the class and ASR tool likely caused some notable degree of improvement in pronunciation for students with mid-level pronunciation. Furthermore, the articulation rate of mid-level students also improved significantly, albeit slightly, which is likely due to the fact that some of the pronunciation lessons (e.g., linking and reduction and rhythm) were more likely to help with fluency rather than pronunciation accuracy. It should be noted that the speech rate of mid-level learners did not improve because they exhibited more pauses. This is likely due to the fact that they became more aware of their pronunciation and were being careful to be more accurate in their speech. As predicted by Skehan's (1998) Limited Capacity Hypothesis, as mid-level learners focused on accuracy, their other aspects of speech, namely fluency, suffered temporary set-backs. However, the increase in articulation rate was enough to prevent speech rates from increasing a significant amount.

Conversely, high-level speakers (those who began with at least 95% intelligibility) did not show any benefits to pronunciation accuracy, likely due to a natural plateau, i.e., it is difficult to show much improvement when many students in the group began with 99% accuracy. However, these students did exhibit an increase in the number of pauses and as a result, a significantly lower speech rate. Furthermore, these students were still very positive towards NatTos. Therefore, students who obtain a 95% intelligibility rate would likely benefit from more advanced pronunciation lessons, such as linking and rhythm, or other speaking activities that focus on improving their fluency and complexity.

With regards to the specific effects of any single class, it seems that that students overall improved on the first lesson (consonant sounds) from the beginning to the end of the lesson, as judged by the NatTos ASR tool. Furthermore, many students reported in the survey that they felt that NatTos was helpful in improving their consonant and vowel pronunciation. However, the only improvement from a single lesson that seemed to have a statistically significant impact on students' overall gains in intelligibility was lesson four (linking and reduction). Here, it should be noted that the results in this area are still a bit unclear for a number of reasons. For example, the number of participants for whom full data is available changed from lesson to lesson, making the data problematic to analyze statistically in a complex model. Furthermore, the pre- and post-lesson ratings conducted via NatTos were of both single words and sentences that were related to the day's pronunciation point, but some of the words and sentences used might have been inappropriate for the NatTos ASR tool, as evidenced in part by the student surveys and detailed below. Finally, since proper pronunciation is a practiced skill, it is possible that one lesson simply was not enough time to impact students to the point that they made a noticeable change, and such changes may have only occurred later after much practice. However, despite the difficulty in analyzing how much of an impact each particular lesson had on students' pronunciation ability, it should be noted that students gained a lot of speaking practice, as evidenced by their repeated attempts at each of the drills, but in and outside the classroom, and also by student surveys which generally viewed the tool positively (both fun and effective).

Since NatTos utilizes an ASR that makes predictions based on N-grams and the probability of certain words appearing singularly or in tandem, as well as based on pronunciation, there were some problems with the tool's ability to check some words and sentences. Specifically, some words were too rare in comparison to other words with similar pronunciations, and thus when used as minimal pairs, one word was always judged as being correct and the other was always judged as being incorrect, regardless of whether or not the target sound was produced correctly. For example, during the second lesson, *nut* and *not* were both used as single words as students were being trained to produce the *ɒ* and *ʌ* sounds. However, as *not* is a very commonly used word, whereas *nut* is only somewhat common, both *nɒt* and *nʌt* were registered as *not* by the ASR, even when pronounced by a native speaker. This problem did not occur with some minimal pairs such as *climb* and *crime*, and therefore, future classes and studies must be careful when creating words for use with ASR pronunciation training tools. A similar problem occurred with sentences, particularly when a strange or unpredicted word was used in a sentence (the ASR would mark correct responses incorrect) or when a very standard sentence was utilized (the ASR would sometimes mark incorrect pronunciations as correct). Finally,

the ASR tool was unable to recognize that homonyms should be marked as correct and would give inappropriate feedback when guessed by the ASR. For example, when practicing with consonant sounds in lesson 1, the ASR guessed *sync* when students were trying to pronounce *sink*. Since these are homophones, full credit should have been awarded, but instead, students were told that their pronunciation was incorrect. Therefore, it is unclear how much the pre- and post-lesson ASR rating given in this study can really be trusted. Future studies should work to pinpoint exact sentences and words that the ASR tool can correctly judge and that are correlated with higher subjective pronunciation ratings by native speakers.

Finally, as indicated by the survey results, future studies should also check which specific sounds are the most problematic for L1 Japanese EFL learners' intelligibility. Though previous studies suggest that sounds such as *r* and *l* are problematic, they tend to focus on these sounds as judged individually by native speakers (e.g., Doleman & Spring, 2014) or individually in terms of phonation (e.g., Goto, 1971) rather than on how much they affect intelligibility. Furthermore, few studies have been able to pinpoint which other specific sounds are difficult and in which contexts. Therefore, future studies that make clear the most problematic pronunciation points in terms of intelligibility could be helpful for planning more effective EFL pronunciation lessons in the future.

## VI. CONCLUSION

The results of this study suggest that the ASR pronunciation training tool NatTos is applicable in online EFL courses, that students receive it well, and that mid-level L1 Japanese EFL learners can objectively improve their intelligibility and articulation rate through pronunciation lessons and practice using it. However, students who already have over 95% accuracy when dictating to an ASR software may exhibit enough single-sound intelligibility and should instead focus on more difficult pronunciation lessons, such as rhythm, or other aspects of speaking. The results of the multiple regression analysis were somewhat unclear, so no hard conclusion regarding which lessons were the most impactful could be reached, but according to the results of the survey, students seemed to think that NatTos was most helpful for consonant and vowel sound lessons. In the future, data should be taken to determine which words and sentences would still make for good practice for the students, but also be predictive of native speakers' judgements of their pronunciation. Based on this data, future studies should also work to determine which sounds and lessons to focus on when teaching EFL pronunciation.

Finally, regarding the pedagogical uses of NatTos and similar ASR programs, we believe that they can be instilled in EFL settings in a number of ways. First of all, the fact that it provides real-time data to teachers regarding which sounds students are having difficulty with makes it ideal for monitoring students during class, regardless of whether it is being done online or in person. Secondly, NatTos being available online and accessible with a number of devices from smartphones to computers allows teachers to potentially give such drills as homework, opening the door to allow teachers to assign speaking homework–something that has thus far been nearly impossible to monitor. Furthermore, NatTos could be used outside of standard speaking classes to provide a more balanced four-skills type of training. For example, in a reading or vocabulary class, teachers could select difficult to pronounce vocabulary words, set them as pronunciation drills in NatTos and then have students not only read the words, but also listen to them and have practice saying them as well–either inside of or outside of the classroom. Finally, receiving a rating on intelligibility can be seen as incredibly motivating for students, as it gives them a clear goal to strive for and encourages them to attempt the drills and speaking assignments repeatedly. For these reasons, we see it as having great potential, as it can be used in a variety of settings and proficiency levels.

## 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

Boersma, P., & Weenink, D. (2019). *Praat: Doing phonetics by computer.* Retrieved from http://www.praat.org

De John, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods, 41*(2), 385-390. https://doi.org/10.3758/BRM.41.2.385

Dolman, M., & Spring, R. (2014). To what extent does musical aptitude influence foreign language pronunciation skills? A multi-factorial analysis of Japanese learners of English. *World Journal of English Language, 4*(4), 1-11. https://doi.org/10.5430/wjel.v4n4p1

Doremalen, J. V., Boves, L., Colpaert, J., Cucchiarini, C., & Strik, H. (2015). Evaluating automatic speech recognition-based language learning systems: A case study. *Computer Assisted Language Learning, 29*(4), 1-19. https://doi.org/10.1080/09588221.2016.1167090

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. Retrieved from 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* (Unpublished master's thesis). Iowa State University, Ames, IA.

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

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

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, September). 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). Ames, IA: Iowa State University.

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

Rogers, H. (2000). *The sounds of language: An introduction to phonetics.* New York, NY: Routledge. https://doi.org/10.4324/9781315838731

Saito, K. (2011). Identifying problematic segmental features to acquire comprehensible pronunciation in EFL settings: The case of Japanese learners of English. *RELC Journal, 42*(3), 363-378. https://doi.org/10.1177/0033688211420275

Saito, Y., & Saito, K. (2016). Differential effects of instruction on the development of second language comprehensibility, word stress, rhythm, and intonation: The case of inexperience Japanese EFL learners. *Language Teaching Research, 21*(5), 589-608. https://doi.org/10.1177/1362168816643111

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. (1998). *A cognitive approach to language learning.* Oxford: Oxford University Press.

Spring, R. (2020). Using multimedia tools to objectively rate the pronunciation of L1 Japanese EFL learners. *ATEM Journal, 25*, 113-124.

Takebayashi, S. (Ed.). (2020). *Pathways to academic English.* Sendai, Japan: Tohoku University.

Tsutomu, A. (2000). *Japanese phonology: A functional approach.* München: Lincom Europa.

Yashima, T. (2002). Willingness to communicate in a second language: The Japanese EFL context. *The Modern Language Journal, 86*(1), 54-66. https://doi.org/10.1111/1540-4781.00136

## APPENDIX

Survey Questions

Survey used in this study (translated from Japanese):

1. Did you enjoy using the NatTos pronunciation training tool?

5 – Yes, very much; 4 – Yes, to some degree; 3 – somewhat; 2 – No, to some degree; 1 – No, not at all

2. Do you think using the NatTos pronunciation training tool helped you to improve your pronunciation?

5 – Yes, very much; 4 – Yes, to some degree; 3 – somewhat; 2 – No, to some degree; 1 – No, not at all

3. What lessons were helpful for you in improving your pronunciation? (check all that apply)

Consonant sounds; Vowel sounds; Stress; Linking and reduction; Rhythm; None of them were helpful

4. Please let us know if you have any other opinions about the pronunciation lessons or the pronunciation training tool used in this class.