Chen, W., Inceoglu, S., & Lim, H. (2020). Using ASR to improve Taiwanese EFL learners' pronunciation: Learning outcomes and learners' perceptions. In O. Kang, S. Staples, K. Yaw, & K. Hirschi (Eds.), *Proceedings of the 11th Pronunciation in Second Language Learning and Teaching conference*, ISSN 2380-9566, Northern Arizona University, September 2019 (pp. 37–48). Ames, IA: Iowa State University.

USING ASR TO IMPROVE TAIWANESE EFL LEARNERS' PRONUNCIATION: LEARNING OUTCOMES AND LEARNERS' PERCEPTIONS

Wen-Hsin Chen, National Taipei University of Technology
Solène Inceoglu, Australian National University
Hyojung Lim, Kwangwoon University

Previous research has shown positive effects of Automatic Speech Recognition (ASR) on the development of second language (L2) pronunciation (Liakin et al., 2017) and on learners' self-perception of their intelligibility (Mroz, 2018). However, empirical studies on the benefits of ASR for pronunciation improvement and learners' engagement remain scant. The present study therefore investigates the effect of ASR on segmental development and explores learners' perception of ASR as a learning tool. A total of 47 Taiwanese university EFL students participated in six sessions of autonomous ASR practice over three weeks; for each practice session, participants used the ASR programs on their mobile phones to record a short reading passage and minimal pairs. Vowel duration and F1 and F2 formant frequency values for minimal pair tokens produced at pretest and posttest were analyzed. The results of the acoustic analysis revealed that learners significantly improved the /æ/-/ɛ/distinction in production, while continuing to struggle with the /i/-/ɪ/ contrast. The participants also reported less positive experience of and attitude toward the use of ASR for pronunciation practice than observed in previous studies. Possible reasons for the findings are discussed.

INTRODUCTION

Automatic Speech Recognition (ASR) is an attractive pedagogical tool to enhance speaking practice, given that it provides low-anxiety learning environments and immediate feedback. ASR has been extensively incorporated into computer-based dictation programs (e.g., Windows Speech Recognition) as well as in commercially available language learning programs (e.g., Rosetta Stone, Tell me More). The increased use of mobile devices allows learners even easier access to ASR; its advantages of being ubiquitous, portable, and personal seem to further appeal to language teachers and learners. In an attempt to investigate the quality of current ASR technology—whether its understanding of non-native speakers' pronunciation is good enough to provide useful corrective feedback—McCrocklin et al. (2019) compared *Google Web Speech (Google)* and *Windows Speech Recognition (WSR)* and concluded that *Google* has considerably improved its accuracy for non-native speakers, surpassing *WSR*.

In addition, past studies have shed light on the positive effect of ASR technology on learners' affective as well as cognitive factors. ASR technology can help raise learners' awareness of their intelligibility (Mroz, 2018) and increase learners' autonomy (McCrocklin, 2016). For instance, McCrocklin found that the university ESL students who were exposed to ASR-based dictation software reported feeling more empowered to practice their pronunciation autonomously. In addition, studies on learner experience of ASR as speaking practice report positive attitudes (Ahn & Lee, 2016; Guskaroska, 2019; Wang & Young, 2015).

The present study explores the effect of ASR technology on the actual development of learners' pronunciation as well as their attitudes. The contribution of ASR applications to learners' oral ability has been illustrated from varying perspectives. Chiu et al. (2007) reported that Taiwanese

freshmen students gained the knowledge of speech acts in English as a result of oral activities with an ASR application. Elimat and AbuSeileek (2014) stated that ASR training helped EFL third graders with their English pronunciation, including producing minimal pairs and sentences. Among the various dimensions of pronunciation constructs (e.g., fluency, intelligibility), we focused on segmental development, examining whether ASR training made any changes to the production of English front vowels among Taiwanese university EFL learners. In addition, because large language classes are the norm in Taiwan and many other EFL settings, results of the current study may have strong pedagogical implications in terms of pronunciation teaching, classroom management, and student learning experience.

Effects of ASR on Pronunciation Development

Despite growing interest in the potential of ASR technology for second language acquisition, the number of studies that have investigated how ASR programs can affect the development of L2 pronunciation, and more particularly of segmentals, is still limited.

To our knowledge, the first study in this area was conducted by Liakin et al. (2015) on the acquisition of the French /y/. They compared three groups of beginner learners of French who either completed weekly pronunciation activities using an ASR application, completed the same activities but with feedback from a teacher, or met with a teacher to practice conversation skills but received no feedback on their pronunciation. At posttest, results revealed that only the ASR group improved their production (but not perception) of /y/.

In another study, Guskaroska (2019) investigated the production of 30 minimal pairs with /i/-/ɪ/; /æ/-/ɛ/; /u/-/ʊ/; /ɑ/-/ʌ/ by 21 Macedonian EFL learners divided between an ASR training group who practiced their pronunciation 20 to 30 minutes a day and a control group. After a period of two weeks, L1 raters' judgments of accuracy scores on learners' pronunciation showed that the ASR group improved significantly more than the control group for /u/, /æ/, and /ʌ/. There was, however, no change for /i/ and /ɛ/.

Finally, another recent study (McCrocklin, 2019) compared the effectiveness of a three-week pronunciation workshop in an ESL listening course where one group of students received face-to-face training and the other "hybrid" group received 50% of face-to-face training and 50% of practice with a dictation program. The workshop targeted vowel contrasts (i.e., /ɛ/-/æ/, /ɑ/-/ʌ/, and /i/-/ɪ/) and consonants (i.e., /ɹ/, /θ/, /ð/, /ʒ/, and /dʒ/). Results of L1 speaker ratings revealed that both groups improved equally from pretest to posttest, with high improvement for /ɪ/ and /ɛ/ (about 10% accuracy gains) and /æ/ (+5.1%), but a decrease for /i/ (-5.7%). Some of these results appear to be in contradiction with Guskaroska's findings and could be due to differences in the participants' L1 and proficiency levels, amount and type of ASR practice, or rating methods.
In sum, although the results of previous studies have revealed some positive effects of ASR practice on pronunciation development, they have also provided conflicting results. Given the very limited number of ASR studies to date, the goal of the current study was to further explore the usefulness of ASR dictation systems on L2 pronunciation learning.

Research Questions

The current study was motivated by the following two research questions:

1. What are the effects of ASR training on segmental accuracy development?
2. What are the attitudes of Taiwanese EFL learners towards English pronunciation practice with ASR?

METHODS

Participants

A total of 49 Taiwanese lower-intermediate EFL learners (18 female) participated in the study. They were between 18 and 20 years old (M = 19.16, SD = 0.51) and had been studying English for six to 13 years (M = 10.06, SD = 2.05). The data were collected in an intact English language course that met 2.5 hours a week for 18 weeks. The focus of the course was on reading and listening skills, with a minor focus on speaking skills albeit with limited attention to pronunciation. In addition to midterm and final exams assessing vocabulary, grammar, listening, and reading skills, a final project involved making short videos introducing tourist attractions in Taipei. The course was taught by an L1 Taiwanese instructor, and students majored mostly in Engineering and Management. Four native speakers of English (2 males, 2 females) served as the baseline.

Materials and Procedures

This study is part of a larger project investigating the use and effects of ASR on the development of speaking skills. All participants were assessed on their pronunciation at pretest and posttest on three types of tasks: 1) a reading passage, 2) a picture description task, and 3) a list of 14 minimal pairs targeting the English vowels /i/-/ɪ/ and /ɛ/-/æ/ and following the model "I said bit, I said beat" (see Table 1). For the pretest and posttest recordings, participants met individually with the lead author or a research assistant in a quiet research room, and completion of the tasks took about six to 10 minutes. In this paper, only the productions of six minimal pairs are analyzed and reported.

Table 1
Minimal pairs for the pretest and posttest. Words analyzed in the study are marked by an asterisk

/i/-/ɪ/	/ɛ/-/æ/
I said beat*	I said said*
I said bit*	I said sad*
I said leave	I said dead*
I said live	I said dad*
I said peel	I said bed*
I said pill	I said bad*
I said sheep*	I said guess
I said ship*	I said gas
I said feel	I said leg
I said fill	I said lag
I said heal	I said pen
I said hill	I said pan
I said peak*	I said ten
I said pick*	I said tan

On the first day, participants also completed a pronunciation attitude questionnaire adapted from Elliott (1995), a motivation questionnaire (Papi & Teimouri, 2012), and a language background questionnaire that included questions regarding participants' experience with ASR programs. At posttest, participants filled out a survey questionnaire on their experiences with ASR. All the questions were presented in the participants' L1 (i.e., Mandarin Chinese).

ASR Training

Participants took part in six sessions of ASR practice over a period of three weeks. The training consisted of three types of production tasks: 1) a short reading passage (average length of 104 words), 2) sentences with minimal pairs targeting both difficult consonants and vowels, and 3) a word list focusing on the minimal pairs /i/-/ɪ/ and /ɛ/-/æ/. Each training session contained novel texts and stimuli that were not found in other training sessions or the pre/posttest. The practice was done outside of class autonomously. However, to ensure that participants were completing the tasks and doing so correctly, participants were asked to record each ASR training session using a video screen-capture and email each video file to their instructor. This allowed the researchers to hear participants' productions and see the written output produced by the ASR system. In a few cases, participants struggled with technological problems and completed their ASR practices on a computer while video-recording their sessions with their mobile phones.

Analysis

For this part of the project, L2 learners' productions were analyzed acoustically in order to explore how ASR pronunciation practice affects the development of two sets of vowel contrasts at a fine-grained level. The pretest and posttest recordings led to a total of 1372 tokens. However, in order to avoid the confounding influence of a liquid or nasal consonant and because some words were clearly problematic for participants (e.g., gas), the results below are based on the analysis of six minimal pairs (see Table 1). After removing tokens with poor audio quality, the total number of

tokens used for the analysis was 588. F1 and F2 measurements were manually extracted from the target words at mid-point along with vowel duration using Praat (Boersma & Weenink, 2018). The analysis of the L1 speaker baseline is reported in Table 2.

Table 2
Mean acoustic values of the English native speakers' vowels. Standard deviations are in parentheses

	/i/	/ɪ/	/ɛ/	/æ/
F1	316.25 (38.83)	464.25 (75.62)	620.33 (89.55)	755.25 (85.78)
F2	2494.25 (347.42)	2034.16 (162.55)	1857.58 (176.27)	1716.83 (296.44)
Duration	103.16 (3.12)	80.91 (3.08)	123.92 (3.75)	105.42 (3.65)

RESULTS

Vowel Production

The first goal of the study was to explore changes in vowel production before and after ASR training. A series of paired-sample *t*-tests were conducted on each vowel separately.

For /i/ (Figure 1), results showed no significant changes in F1 values, $t(254) = -1.168$, $p = .244$, and F2 values, $t(254) = -829$, $p = .408$, but significant increase in duration, $t(254) = -2.780$, $p = .006$.

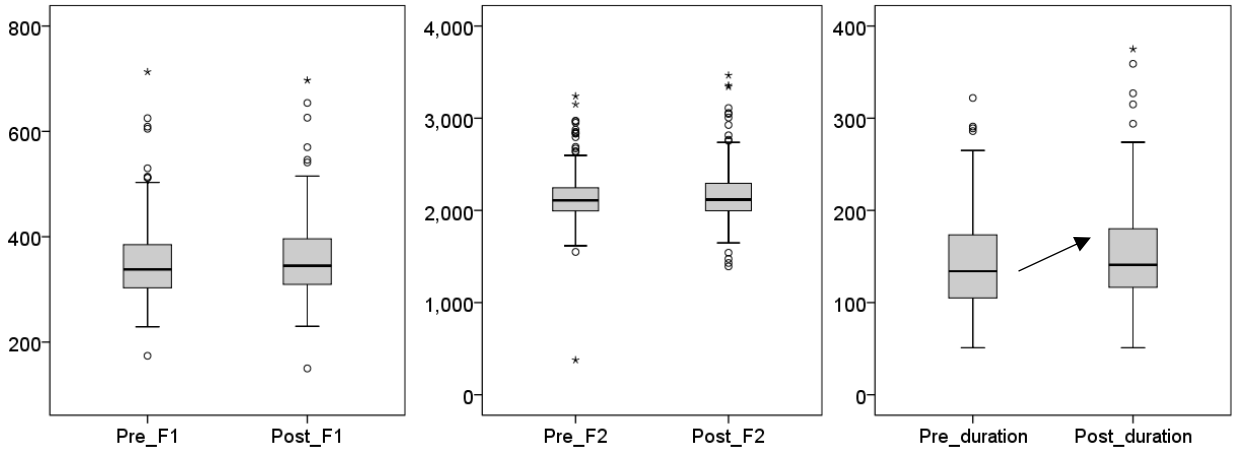

Figure 1. Acoustic measures for the production of /i/ by Taiwanese EFL learners before and after ASR practice.

For /ɪ/, there was a significant increase in F1 values, $t(252) = -3.501$, $p = .001$, but no significant changes in F2 values, $t(252) = -.992$, $p = .322$, and in duration, $t(252) = 1.907$, $p = .058$ (Figure 2).

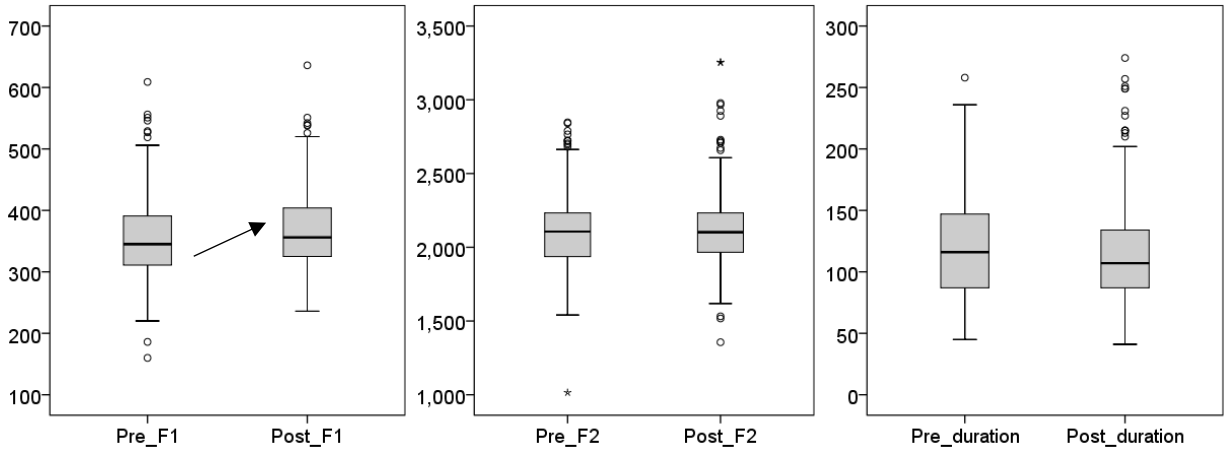

Figure 2. Acoustic measures for the production of /ɪ/ by Taiwanese EFL learners before and after ASR practice.

The results for /ɛ/ revealed a significant increase in F1 values, $t(256) = -5.905$, $p < .001$ and in duration, $t(256) = -3.240$, $p < .001$, but no significant changes in F2 values, $t(256) = -.155$, $p = .877$ (Figure 3).

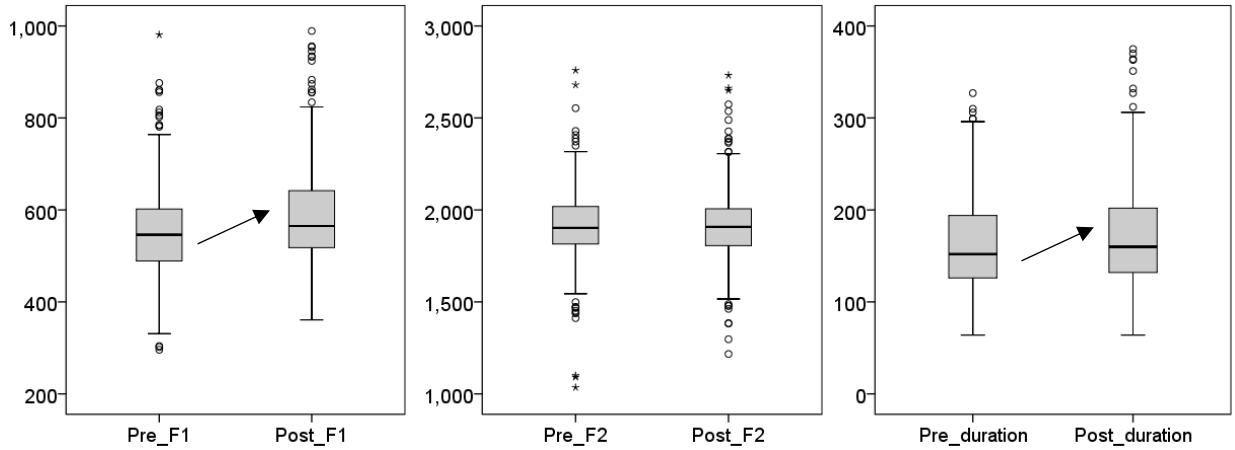

Figure 3. Acoustic measures for the production of /ɛ/ by Taiwanese EFL learners before and after ASR practice.

Finally, for /æ/, there was a significant increase in F1 values, $t(274) = -8.124$, $p < .001$ and in duration, $t(274) = -4.530$, $p < .001$, but no significant changes in F2 values, $t(274) = -.163$, $p = .871$ (Figure 4).

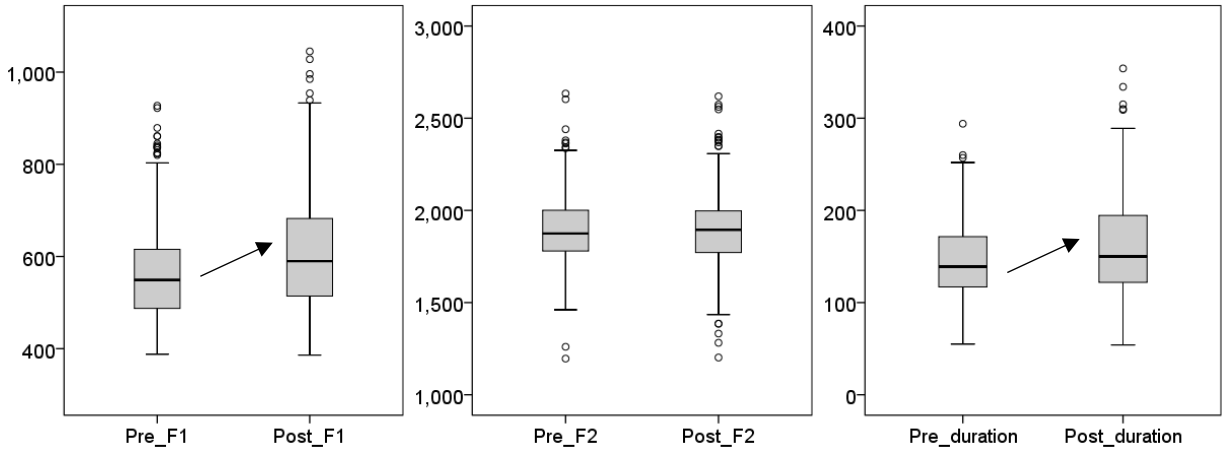

Figure 4. Acoustic measures for the production of /æ/ by Taiwanese EFL learners before and after ASR practice.

Figure 5 illustrates the participants' average vowel space at pretest and posttest, along with a reference to native speakers' average norms. The data clearly show that among the L2 learners /i/ and /ɪ/ are merged, with a Euclidean distance becoming even closer at posttest. Also represented on the figure is the single category /æ/-/ɛ/ at pretest that learners start to distinguish at posttest.

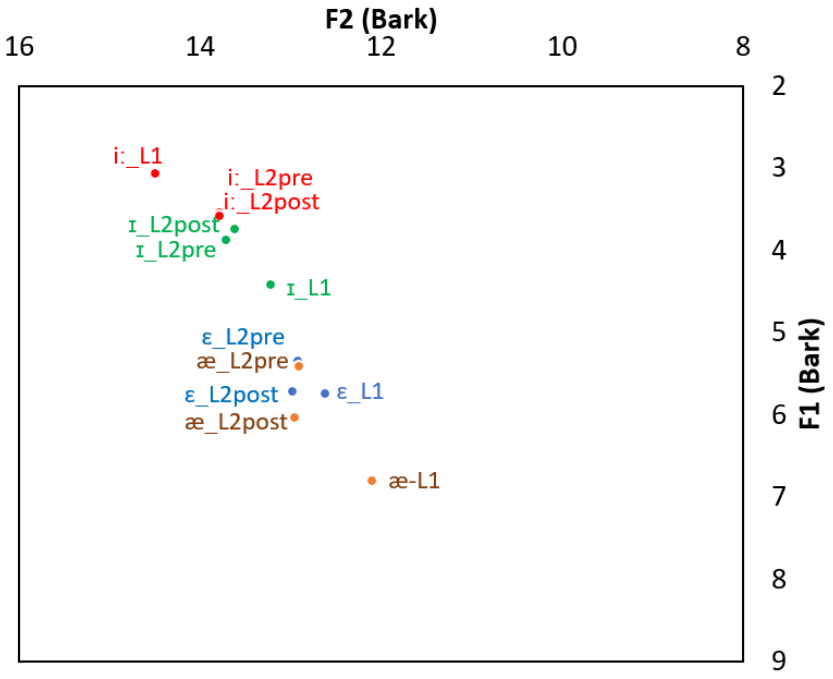

Figure 5. Average formant values of /i/, /ɪ/, /ɛ/ and /æ/ by L1 Taiwanese speakers at pretest and posttest, and L1 English norms.

Learners' Perception and Experiences with ASR

The second goal of this study was to explore students' perception of ASR. Their experiences using ASR and their attitudes towards this technology as a practical tool to practice L2 English pronunciation were collected via a set of questions in the exit questionnaire.

The first question focused on how well ASR recognized what participants said when they spoke English during the six sessions of autonomous practice. Figure 6 shows a large range of experiences with a higher number of the participants reporting average recognition of their speech by the ASR program. There were, however, more participants reporting great success of ASR results than participants who appeared to have experienced very poor performance.

When further prompted to provide reasons why ASR worked well for them or did not work, three (out of 49) participants cited their pronunciation as the sole issue, 11 named the current state of ASR as the source of the problem, and 35 thought that it was a combination of the two.

Another question prompted the participants to rank the usefulness of ASR as a practice tool to help improve their pronunciation in English based on their recent experience. Results (Figure 7) showed an average of 5.5 (out of 10), with no participant finding ASR very useful to practice pronunciation and a few reporting very negative opinions of its usefulness.

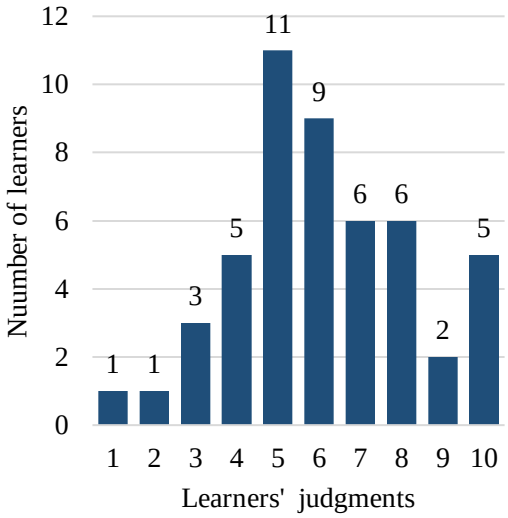

Figure 6. Learners' report of how well ASR recognized their speech from 1 (very poor) to 10 (excellent).

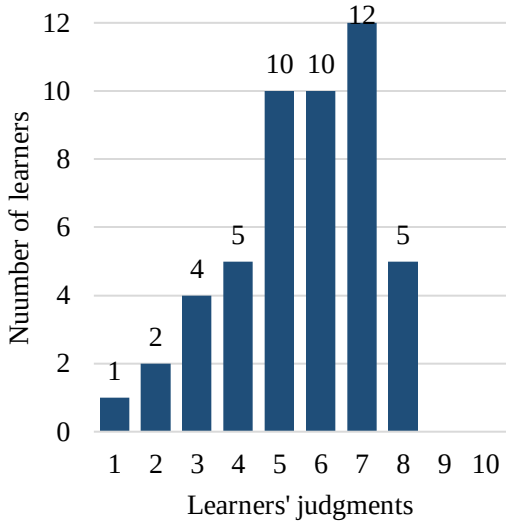

Figure 7. Learners' beliefs about the usefulness of using ASR to practice English pronunciation from 1 (not useful at all) to 10 (extremely useful).

The posttest survey also invited participants to report the most common problems they encountered while using ASR in English. Several themes emerged from the participants' free responses. Below, we identify the main issues along with examples of participants' quotes translated from Chinese:

- ASR has problems recognizing minimal pairs (*n* = 15)

"When I read words that sound very similar, no matter how many times I tried, ASR always showed the same words."
"Often, ASR failed to recognize my long and short vowels."

- Easier for ASR to recognize sentences than isolated words (*n* = 11)

"ASR knows what you say according to the context. Compared to words, ASR is better at recognizing sentences and passages."
"ASR will more accurately recognize what you say when you read a complete sentence. The more context you provide, the more accurate it is."

- ASR does not recognize what I say (*n* = 10)

"Basically ASR never accurately recognized what I said."

- Limitations of the equipment used (*n* = 8)

"Sometimes I turned on ASR and then read a sentence, but it just didn't do any voice typing."
"If I read quickly or read several sentences in a row, ASR always failed to recognize what I said. I had to read slowly so that ASR could recognize my pronunciation."

In addition to the above comments, five participants reported having encountered no problem with ASR during their six sessions of practice.

Particularly relevant from a pedagogical viewpoint, a last question asked participants whether they intended to continue using ASR in the future to practice English pronunciation. Surprisingly, only 22% of the participants responded that they would. These participants elaborated in a follow-up question that ASR made them aware of their pronunciation shortcomings, stating, for example *"I realized my pronunciation was not good. I think I made some improvements"* and *"Previously I thought that my pronunciation of some common words was good. After using ASR, I realized my pronunciation was not good."*

Participants who reported not planning to use ASR as a way to work on their English pronunciation had different rationales. Some seemed to struggle with the repeated trials needed for ASR to recognize their speech, highlighting a time issue, *"If ASR can't accurately recognize what I say, I will have to repeat several times. This is a waste of time"* and *"ASR is for people with patience."* Others stated that ASR technology is not advanced enough for them to use it as a pedagogical tool, *"ASR technology hasn't fully developed yet"* or expressed self-conscious concerns *"I feel it's silly to talk to my phone."*

Finally, a couple of participants appeared not to have understood the pedagogical reason behind using ASR as a way to practice their pronunciation. Some seemed to perceive ASR as an end rather than as a means to develop their speaking skills, with a learner writing "*I will use ASR after I improve my pronunciation. After all, speaking is faster than typing*." This is in direct contradiction with another participant who was also not interested in using this technology for English pronunciation practice and who responded that *"typing is faster than the voice typing function of ASR."* This discrepancy between what learners can do as an autonomous language learning activity

and what cell phone users do in real life situations is represented even further in a participant's comment that *"it is more convenient to type on my phone because I won't be able to talk in some situations."*

DISCUSSION

The first goal of the study was to explore the effects of autonomous ASR practice on the development of four English vowels. Overall, the results indicated that ASR practice was only beneficial for the /ɛ/-/æ/ pair, with acoustic features closer to native-like features in terms of height but not backness. Most importantly, learners appeared to make greater distinction between the two vowels which were not distinguished in production at pretest. In regards to /i/ and /ɪ/, a vowel contrast that is difficult to acquire by L1 Chinese speakers (Han, 2013; Zhang & Yin, 2009), the results showed that they were produced as one category at both pretest and posttest. This is consistent with Flege's (1995) Speech Learning Model that the phones that do not contrast in the L1 are difficult to perceive for L2 learners, which will be reflected in the learner's production of the contrast.

In relation to previous research, the results of the current study appear to differ from McCrocklin (2019) who found high improvement for /ɪ/ and a decrease for /i/ (-5.7% accuracy), but are consistent with a noted improvement for /ɛ/ (+10% accuracy) and /æ/ (+5.1% accuracy). Despite the difference in participants' L1 background, our findings also aligned with Guskaroska's (2019) report that ASR practice did not lead to significant changes for /i/, but helped improve the production of /æ/. Interestingly, that study did not see a benefit of ASR pronunciation practice for /ɛ/. Due to the limited number of ASR studies on the development of segmentals, firm conclusions cannot be reached. Yet, participants in the three studies appeared to have benefited from ASR practice with regards to the /æ/-/ɛ/ contrast. This will need to be confirmed in future studies. Moreover, the current study examined learners' production acoustically, and it would be worth conducting further analysis with L1 ratings. This comparison might also shed light on the relationship between measurable acoustic features of speech and listeners' ratings of L2 learners' vowels, and how (if at all) these small improvements in the vowel space are perceived by native listeners.

In terms of learner experience and attitude, the results of the current study appear somewhat less positive than previous studies (Ahn & Lee, 2016; Guskaroska, 2019; Liakin et al., 2017; McCrocklin, 2016; Mroz, 2018; Wang & Young, 2015). This is particularly apparent with regards to the fact that 78% of the participants had little inclination to continue using ASR to practice their pronunciation. Several explanations can be invoked for these findings. First, differences in the participants' profiles (e.g., age, majors) and learning settings (e.g., immersion, foreign language setting) could have influenced the results. To date, the paucity of ASR research does not permit generalization. The data in the current study were collected from low-intermediate learners enrolled in a technological university in Taiwan, whereas previous research was conducted with learners of French in second language (Liakin et al., 2017) and foreign language (Mroz, 2018) settings, middle school students in South Korea (Ahn & Lee, 2016), EFL learners in Macedonia (Guskaroska, 2019), and ESL learners (McCrocklin, 2016).

Related to the difference in institutional setting comes the issue of motivation, autonomy, and access to the target language. The comparison between the current study and McCrocklin (2016)

is particularly interesting due to the similar linguistic background of the participants (i.e., mostly L1 Chinese) and the different settings. McCrocklin's (2016) three-week ASR pronunciation workshop was fully embedded in an ESL listening course taught on a US Midwestern campus, whereas participants in the present study completed six sessions of ASR practice outside of class. Although they were required to email the videos of their practice sessions and received marks for their participation, the lack of connection between the course and ASR practice might have contributed to less positive attitudes. In addition, not having access to a (native) model to demonstrate the accuracy of ASR probably reinforced participants' skepticism toward the current capabilities of this technology. Indeed, some of the comments pointed to a belief that ASR programs are not developed enough, suggesting that some learners did not seem to consider their pronunciation as poor and the source of the problem.

CONCLUSION AND FUTURE DIRECTIONS

While the current study provided insights into the effectiveness of ASR dictation practice on segmental accuracy and on EFL learners' perceptions of ASR as a pedagogical tool, there are some limitations that need to be addressed. First, the analysis only focused on the participants' pretest and posttest oral productions, and the accuracy of feedback provided during the ASR training sessions was beyond the scope of the study. Future analyses will focus on the mobile-phone screen recordings—capturing both the ASR output and the participants' voices—to examine how accurate mobile-based ASR technology is at providing feedback. Second, the contribution of individual differences should also be investigated, especially with regards to the relationship between pronunciation development and motivation and attitudes towards ASR. The data presented in the current study captured high variability in learners' performance and perception of ASR, and more research is needed on this topic.

ABOUT THE AUTHORS

Wen-Hsin Chen is an Assistant Professor in the Media Center at National Taipei University of Technology, Taipei, Taiwan. Her research interests include second language acquisition, language classroom interaction, language processing, and Chinese language and culture. Contact information: wchen33@ntut.edu.tw

Solène Inceoglu is a Lecturer (Assistant Professor) in French in the School of Literature, Languages, and Linguistics at the Australian National University. She received her Ph.D. in Second Language Studies from Michigan State University. Her research focuses on second language acquisition, second language speech perception/production, pronunciation instruction, and psycholinguistics. Contact information: solene.inceoglu@anu.edu.au

Hyojung Lim is an Assistant Professor in the Department of English and Industry at Kwangwoon University, Seoul, Korea. Her research revolves around psycholinguistics, language testing, and individual differences. Contact information: lim@kw.ac.kr

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*, 778–786.

Boersma, P., & Weenink, D. (2018). Praat: doing phonetics by computer (Version 6.0.42) [Computer software]. Retrieved from http://www.praat.org.

Elliott, A. R. (1995). Foreign language phonology: Field independence, attitude, and the success of formal instruction in Spanish pronunciation. *The Modern Language Journal*, *79*, 530–542.

Elimat, A. K., & AbuSeileek, A. F. (2014). Automatic speech recognition technology as an effective means for teaching pronunciation. *JALT CALL Journal, 10*(1), 21-47.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 233–277). Timonium, MD: York Press.

Guskaroska, A. (2019). *ASR as a tool for providing feedback for vowel pronunciation practice*. Unpublished master thesis, Iowa State University, Ames.

Han, F. (2013). Pronunciation problems of Chinese learners of English. *ORTESOL Journal, 30*, 26-30.

Liakin, D., Cardoso, W., & Liakina, N. (2015). Learning L2 pronunciation with a mobile speech recognizer: French /y/. *CALICO Journal*, *32*, 1–25.

Liakin, D., Cardoso, W., & Liakina, N. (2017). Mobilizing instruction in a second-language context: Learners' perceptions of two speech technologies. *Languages, 2*, 1-21.

McCrocklin, S. (2016). Pronunciation learner autonomy: The potential of Automatic Speech Recognition. *System*, *57*, 25–42.

McCrocklin, S. (2019). ASR-based dictation practice for second language pronunciation improvement. *Journal of Second Language Pronunciation, 5*, 98–118.

McCrocklin, S., Humaidan, A., & Edalatishams, I. (2019). ASR dictation program accuracy: Have current programs improved? In *Proceedings of the 10th Pronunciation in Second Language Learning and Teaching Conference* (pp. 191–200). Ames, IA: Iowa State University.

Mroz, A. (2018). Seeing how people hear you: French learners experiencing intelligibility through automatic speech recognition. *Foreign Language Annals*, *51*, 617–637.

Papi, M., & Teimouri, Y. (2012). Dynamics of selves and motivation: A cross-sectional study in the EFL context of Iran. *International Journal of Applied Linguistics, 22*(3), 287-309.

Chiu, T. L., Liou, H. C., & Yeh, Y. (2007). A study of web-based oral activities enhanced by Automatic Speech Recognition for EFL college learning. *Computer Assisted Language Learning, 20*(3), 209-233.

Wang, Y. H., & Young, S. S. C. (2015). Effectiveness of feedback for enhancing English pronunciation in an ASR-based CALL system. *Journal of Computer Assisted Learning*, *31*, 493–504.

Zhang, F. C., & Yin, P. P. (2009). A study of pronunciation problems of English learners in China. *Asian Social Science, 5*(6), 141-146.