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, 36(5–6), 861–884.

This is an Accepted Manuscript of an article published by Taylor & Francis in Computer Assisted Language Learning on 26 Jul 2021 (published online), available at: http://www.tandfonline.com/10.1080/09588221.2021.1952272.

Mobile-assisted pronunciation learning with feedback from peers and/or automatic speech recognition: A mixed-methods study

Abstract

Although social networking apps and dictation-based automatic speech recognition (ASR) are now widely available in mobile phones, relatively little is known about whether and how these technological affordances can contribute to EFL pronunciation learning. The purpose of this study is to investigate the effectiveness of feedback from peers and/or ASR in mobile-assisted pronunciation learning. 84 Chinese EFL university students were assigned into three conditions, using WeChat (a multi-purpose mobile app) for autonomous ASR feedback (the Auto-ASR group), peer feedback (the Co-non-ASR group), or peer plus ASR feedback (the Co-ASR group). Quantitative data included the pronunciation pretest, posttest, and delayed posttest, and students' perception questionnaires, while qualitative data included students' interviews. The main findings are: (a) all three groups improved their pronunciation, but the Co-non-ASR and the Co-ASR groups outperformed the Auto-ASR group; (b) the three groups showed no significant difference in perception questionnaires; and (c) the interviews revealed some common and unique *technical*, *social/psychological*, and *educational* affordances and concerns about the three mobile-assisted learning conditions.

Keywords: automatic speech recognition, mixed methods, mobile-assisted language learning, peer feedback, pronunciation

1. Introduction

Pronunciation learning requires immediate, personalized feedback that helps EFL students address the gaps between their performances and the desirable ones (Bodnar et al, 2017; van Doremalen et al., 2016). However, due to limited in-class time and a large class size, not much pedagogical attention is given to timely, personalized feedback (Levis, 2007; Neri et al., 2008). To redress these issues, many technical solutions have been proposed, ranging from speech visualization to automatic speech recognition (ASR), from interacting with chatbots to practicing pronunciation through social networking media (Pennington & Rogerson-Revell, 2019). In this article, we focus on a multi-purpose mobile app, WeChat, in pronunciation learning. We situate our study in two bodies of research: ASR and collaborative pronunciation learning. As WeChat comes with dictation ASR and social networking functions, we want to examine the effectiveness of feedback from peers and/or ASR in mobile-assisted pronunciation learning. In the following sections, we will first review the theoretical background and relevant studies on ASR and collaborative pronunciation learning. We will then report on a mixed-methods study comparing the learning outcomes and perceptions of 84 Chinese EFL students. We will discuss the common and unique (dis)advantages of three learning conditions and offer pedagogical implications for instructors to make informed choices pertinent to using ASR and social apps for pronunciation learning.

2. Theoretical Background

This study is guided by skill acquisition theory (DeKeyser, 2017) and sociocultural theory (Lantolf, 2012). According to skill acquisition theory, deliberate practice is necessary to turn knowledge about a skill into behaviors acting on the knowledge. This transition from "knowledge that" to "knowledge how" is called proceduralization (DeKeyser, 2017). With more practice, the process becomes more effortless and achieves automatization. This is particularly true for pronunciation learning. Students need deliberate practice to turn pronunciation rules into behaviors of producing target pronunciation features. Through repeated practice, production becomes highly effortless and automatic. In this regard, technical applications afford abundant opportunities for students to engage in repeated practice to proceduralize and automatize their pronunciation. In addition to practice, feedback is equally important to achieve proceduralization and automatization (Sato & Lyster, 2012). In technology-assisted pronunciation learning, students can obtain timely feedback from software programs. For instance, a program may display "Well done" to indicate that the pronunciation is acceptable (Neri et al., 2008). This diagnostic information can contribute to skill acquisition.

Another way to obtain diagnostic information is through peer feedback. Peer feedback is an epitome of collaborative learning in that students provide and receive "social support and scaffolding" in group work (Hu & Lam, 2010, p. 373). Underpinned by sociocultural theory, scaffolding refers to the support offered by experts or more capable peers for students to perform tasks that might not be completed by students alone (Lantolf, 2012). In pronunciation learning, students can benefit from more capable peers, who point out pronunciation issues and demonstrate pronunciation methods. As such, peer feedback functions as scaffolding specific to students' learning needs. After students proceduralize and automatize their pronunciation, scaffolding can be removed, as the building metaphor implies. This study draws on the theoretical insights outlined previously and seeks to understand how ASR and peer feedback can contribute to pronunciation learning.

3. Literature review

3.1 ASR and pronunciation learning

ASR has been widely deployed in commercial and open systems to improve learners' pronunciation and/or speaking skills (van Doremalen et al., 2016). Studies have found that ASR contributes to learners' pronunciation quality (Neri et al., 2008), accuracy of spoken grammar structures (de Vries et al., 2015), and mastery of speech acts (Chiu, Liou, & Yeh, 2007). Additionally, ASR affords a relaxed and enjoyable environment for learners (Bodnar et al., 2017) and reduces their speaking anxiety (Bashori et al., 2020). The effectiveness of ASR is highlighted by Golonka et al. (2014), after reviewing 350 empirical studies on language learning technologies.

One recent development of ASR-assisted pronunciation learning is the scholarly interest in *mobile-based dictation* ASR. Different from computer-based ASR systems (e.g.

van Doremalen et al., 2016) and language learning apps (e.g. Loewen et al., 2019), *mobile-based dictation* ASR (e.g. iPhone Siri) does not generate an assessment score or highlight mispronounced syllables. Instead, learners need to examine the dictation text to identify pronunciation issues (Liakin et al., 2015). Although mobile-based dictation ASR feedback seems simple, it contributes to learners' pronunciation learning. For instance, Liakin et al. (2015) compared students' pronunciation learning in three conditions: ASR feedback (provided by an app Nuance Dragon Dictation), teacher feedback, and no feedback. The ASR feedback group improved significantly, while the other two groups did not. More recently, Mroz (2018) found that learners benefited from using Gmail as a dictation ASR tool in read-aloud and semi-spontaneous speaking tasks. The learners improved their motivation and willingness to communicate and over half of them improved their oral proficiency. While these studies point to the benefits of mobile-based dictation ASR, they primarily focus on *autonomous* practice. Learner autonomy is certainly important for language learning, but so are learner interaction and collaboration (Mackey, 2020). Given the enhanced connectivity enabled by mobile technologies, there is a need to understand the extent to which ASR-assisted *collaborative* practice can improve students' pronunciation.

3.2 Collaborative pronunciation learning in virtual environments

In collaborative pronunciation learning, peer interaction affords rich opportunities for learners to practice target pronunciation features with peers' scaffolding. Recent studies have shown that peer feedback in conjunction with computer-based ASR contributes to students' perception and performance in pronunciation learning. For instance, Tsai (2019) found that students enjoyed peer feedback and became more aware of the (de)merits of their pronunciation. In an experiment (Tsai, 2015), English major students were assigned into three treatment conditions (individual computer-assisted vs. collaborative computer-assisted vs. individual non-computer-assisted). All three groups improved their pronunciation, although the improvement did not significantly differ from one another. Evers and Chen (2020) compared two groups of adult learners who either practiced with computer-based ASR plus peer feedback or with ASR feedback alone. They found that the former outperformed the latter in the comprehensibility of read-aloud sentences and in spontaneous conversation. Taken together, these studies have demonstrated that technology-assisted peer interaction is beneficial to learners' affect and performance in pronunciation learning.

3.3 Aims and research questions

A review of related literature shows that previous studies have tended to focus on computer-based ASR feedback and its comparison with face-to-face peer feedback. As dictation ASR is increasingly ubiquitous in mobile phones and students maintain their social network mainly through mobile devices, it is interesting to know whether ASR coupled with the affordances of "mobility" and "peer connectivity" (Godwin-Jones, 2011, p. 7) can lead to better pronunciation learning. As such, our study aims to examine

the effects of mobile-assisted peer feedback and/or ASR feedback on EFL students' pronunciation learning. The study was guided by the following research questions:

1. Do three mobile-assisted conditions (i.e. peer feedback, ASR feedback, and a combination of both) contribute to students' pronunciation learning?
2. Do students under three mobile-assisted conditions differ in their pronunciation learning?
3. How do students perceive mobile-assisted peer and/or ASR feedback?

4. Methodology

4.1 Research context

This study adopted a mixed-methods design (Creswell & Plano Clark, 2007), which involved pronunciation tests and soliciting participants' perceptions through questionnaires and interviews. The tests and questionnaires generated quantitative data, while the interviews provided qualitative data. The rationale behind the research design was twofold. First, it focused on both learning effects and student perceptions, thus allowing a comprehensive account of both outcomes and experiences of the mobile-assisted pronunciation process (Loewen et al., 2019). Second, student perceptions were captured by quantitative and qualitative data, the triangulation of which allowed us to systematically identify, compare, and explain perceived differences among mobile-assisted peer and/or ASR feedback.

4.2 Participants

The study initially involved three parallel classes (each with 60 second-year students) at a first-tier university in China. The male-to-female gender ratio was 1:2 and the average age was 19.8 years old (SD = 0.88). They were enrolled in a *College English* course, a mandatory English enhancement course at the university. As the students came from different majors, their English proficiency levels varied. About 40% could be positioned at the A1 level, 40% A2, and 20% B1 in the Common European Framework of Reference for Languages. One teaching objective of the course was to enhance the students' pronunciation. Due to the COVID-19 pandemic, the course was delivered online and mobile-assisted pronunciation task was designed accordingly.

4.3 Tool

WeChat (a multi-purpose app) was used for the mobile-assisted pronunciation task for five reasons. The first four conformed to Mroz's (2018) criteria (cost, accessibility, familiarity, and technology): WeChat is free, working on multiple platforms (e.g. Android and iOS), widely popular among Chinese mobile-phone users, and empowered by deep learning technology. The fifth reason was specifically important for this study: WeChat could function as a networking tool to synchronously audio-chat with friends and as a speech-to-text recognition tool. We acknowledge that WeChat is not designed for language learning, but it promises multiple affordances germane to language learning (Jin, 2018). For instance, the synchronous audio chat function can be used among classmates for peer feedback not

constrained by physical proximity. This was an important affordance for the students who stayed home but were able to collaborate during the pandemic. Additionally, the ASR function in WeChat (see Appendix 1) is speaker-independent, meaning that "it does not require speaker training prior to use" (Liakin et al., 2015, p. 3) and does not get used to a speaker's accent or mispronunciation.[1] Therefore, it is suitable for pronunciation training, as learners can read the ASR transcription and identify the mispronounced words (Evers & Chen, 2020).

4.4 Procedures

The study was structured in three stages (see Figure 1). In the first stage, all the students took a pretest, which required them to read eight sentences (see Appendix 2). These sentences were selected from the textbook guided by two criteria. First, based on the instructor's 12-year teaching experience at the university, the students had tended to experience difficulties with these eight sentences, although they had been explicitly taught the International Phonetic Alphabet in their middle schools. Second, the eight sentences contained vowels and consonants that were difficult for Chinese speakers of English, such as /əː/ (Chang, 1987), /iː/, /ʌ/, /eɪ/, /ɑː/ (McAndrews & Thomson, 2017), /θ/, /ð/ (Deterding, 2006), /v/ (McAndrews & Thomson, 2017), /tr/ and /str/ (Liaw, 2014). The sentence-reading task was used rather than a spontaneous speaking task for two reasons. First, the sentence-reading task, as a type of controlled production, "allowed us to make direct comparisons within and between individuals" (Saito & Saito, 2017, p. 604). Second, based on a meta-analysis of 77 studies, Saito and Plonsky (2019) found that "there was less measurement error when learner production was controlled as opposed to spontaneous" (p. 679). During an online session, the students were given ten minutes to (a) record their performances and (b) submit the recordings for grading. Two raters rated their performances and the averages were taken as the students' pretest scores (see raters' information and inter-rater reliability in Section 4.5).

[Insert Figure 1 here]

In the second stage, one week later, the three classes were assigned into three conditions. For Class One, the students were paired up based on the rank-order of pretest scores with a constant rank difference (Huisman et al., 2017). For instance, of the 60 students, the top 1 student was paired up with the 31st one, the 2nd student with the 32nd one, and so on. Within a dyad, the higher-grade student acted as a feedback giver, while the lower-grade student acted as a feedback receiver. The dyads used WeChat's synchronous audio chat function to conduct a one-to-one feedback session. This treatment condition was called the Collaborative-non-ASR Group (or the Co-non-ASR Group). For Class Two, the grouping method and the role assignment were exactly the same as those in Class One.

[1] The accuracy of ASR was tested and all words in the eight target sentences could be recognized by the app.

However, prior to their one-to-one feedback session, the feedback receivers recorded the eight sentences in WeChat and sent them to feedback givers. Then, both students in the dyad used WeChat's ASR function and inspected the app-generated transcripts. Building on this, they conducted their one-to-one feedback session via synchronous audio chat and with reference to the ASR transcription/feedback. This treatment was called the Collaborative-ASR Group (or the Co-ASR Group). For Class Three, the students were not grouped but engaged in an autonomous self-practice session. They practiced the eight sentences with WeChat's ASR transcription. They could repeat multiple times until they found the results desirable. This was the Autonomous-ASR Group (or the Auto-ASR Group).

For the Co-non-ASR and the Co-ASR groups, when the sentence-reading task was announced, the instructor explained some collaboration strategies (e.g. identifying feedback foci and consulting sources) to the students. For the Co-ASR and the Auto-ASR groups, the students were shown how to locate potential pronunciation issues and practice pronunciation until attempts were successful (e.g. Step 4 in Appendix 1). Following these instructions, the three classes conducted the sentence-reading task with peer and/or ASR feedback. The expected study effort was about 60 minutes. Immediately after the task, the students recorded the eight sentences again and submitted them for grading (the immediate posttest). They also took an online questionnaire that tapped their fresh memory and solicited the perceived usefulness of using the mobile app.

In the third stage, one week later, the students took an unannounced delayed posttest. They were given ten minutes to (a) record the eight sentences and (b) submit them for grading. Two weeks after the delayed posttest, 18 students were invited and consented to participate in interviews that solicited their retrospective, reflective insights about the mobile-assisted feedback task. The interviewees were randomly selected from the lower-grade students (see the next paragraph), reflecting the gender ratio of the student population (i.e. two males and four females from each of the three groups).

In this study, we decided to focus on the students with lower grades in the pretest because they had more potential/room to improve their pronunciation and thus provided more opportunities to manifest the effect of different treatment conditions.[2] Additionally, the focus on the lower-grade students across the three conditions ensured the between-group comparability (i.e. all feedback receivers). In other words, the comparability would not be ideal if we included the higher-grade students, because they acted as feedback givers in the Co-non-ASR and Co-ASR groups, but as feedback receivers in the Auto-ASR group. As some students did not submit the full set of data, the subsequent analysis was based on a total of 84 lower-grade students (28 students × 3 groups).

4.5 Measures

For the quantitative analysis, the independent variable was the treatment condition (i.e. sources of feedback). The dependent variables consisted of one global construct, *comprehensibility*, and two specific constructs: *segmental accuracy* and *word stress*

[2] For instance, the mean comprehensibility scores of the higher-grade students in the pretest were 7.59 (out of 9), 7.60, and 7.56, respectively in the Co-non-ASR, Co-ASR, and Auto-ASR groups.

accuracy (Saito & Plonsky, 2019; Crowther et al., 2016). To understand students' perceptions about the three mobile-assisted learning conditions, questionnaire and interview data were also collected. Details of measurements and data collection were described in the following paragraphs.

Comprehensibility. This construct measures the ease of understanding when listening to a speech sample. It has been recognized as a useful concept "for assessing improvements in the development of L2 pronunciation" (Dlaska & Krekeler, 2013, p. 30) and a prioritized goal for L2 pronunciation teaching (Levis, 2005). The students' speech samples were rated by two raters. One had a PhD degree in applied linguistics and the other was a PhD student in the same field. Although both raters were Chinese speakers of English, previous studies have found that nonnative and native speakers displayed no differences in ratings (Crowther et al., 2016). The raters participated in a calibration session, in which they practice-rated multiple speech samples and discussed disagreements. Then, the recordings from three tests in three conditions were randomized, anonymized, and rated independently by the raters. The inter-rater reliability was measured by Pearson's correlation ($\alpha = 0.932$). Given the high reliability, the two raters' scores were averaged as the comprehensibility scores for the subsequent analysis.

Segmental Accuracy. Based on a review of literature (Chang, 1987; Deterding, 2006; Liaw, 2014; McAndrews & Thomson, 2017), we decided to focus on five vowels (/iː/, /ʌ/, /eɪ/, /ɑː/, and /əː/) and five consonants/consonant clusters (/θ/, /ð/, /v/, /tr/, and /str/) that pose pronunciation difficulties to Chinese speakers of English. A total of 30 instances of these segmentals were found in the eight sentences (see Appendix 2). The segmental accuracy was coded as 1 (correct) or 0 (incorrect or missing) independently by two research assistants. Discrepancies were resolved through discussion.

Word Stress Accuracy. Of the 21 words that contained a stressed syllable, 10 words were selected as the target assessment units (see Appendix 2). A pilot round of coding all the 21 words in one third of the pretest recordings showed that 11 words were pronounced with a correct stress. Hence, focusing on the 10 words with stress problems could enable us to know the extent to which the students improved after the pronunciation task. In line with the literature (e.g. Crowther et al., 2016), the accuracy of word stress was coded as 1 (correct stress, e.g. FORward) or 0 (misplaced or missing stress, e.g. forWARD) independently by two research assistants. Discrepancies were resolved through discussion.

Questionnaire. The questionnaire consisted of six items (see Table 2). The students rated their perceptions on a 6-point scale (with 1 meaning "totally disagree" and 6 meaning "totally agree"). The items were adapted from Luo (2016).

Interview. We adapted from Mroz's (2018) interview protocol focusing on the students' experiences, perceptions, and suggestions. In the first part, they were asked to recount their experiences of the sentence-reading task. In the second part, they described the advantages and disadvantages of the feedback configurations. In the final part, they offered their suggestions to improve the mobile-assisted learning task. The interviews were conducted in Chinese and lasted between 30 and 60 minutes. The extracts cited in this article were translated by the authors.

4.6 Data analysis

As some of the data were not normally distributed, for the quantitative analysis, non-parametric tests were used. To answer the first research question (within-group improvement), the Wilcoxon Signed Rank tests were used to see whether the delayed posttest outcome measure scores were significantly higher than the pretest scores within each group, which suggested pronunciation improvement (Mroz, 2018). To answer the second research question (between-group difference), Kruskal-Wallis tests were conducted to determine whether the three groups differed significantly in terms of the outcome measures (Crowther et al., 2016). If significant results were found, Mann-Whitney tests were used to locate the sources of differences.

To answer the third research question (student perception), the ratings of the six Likert-scale items were compared across three groups, using Kruskal-Wallis tests and Mann-Whitney tests (when necessary). Additionally, interviews were transcribed and analyzed according to Miles, Huberman and Saldaña's (2014) two-cycle coding procedure. In the first cycle, we independently read the transcripts and assigned "labels to data to summarize in a word or short phrase" the students' attitudes, experiences, and thoughts (Miles et al., p. 74). For instance, two labels "convenience" and "repetition" were assigned to this interview extract: "It was very convenient. We could talk to each other anytime anywhere…We could practice again and again." In the second cycle, we performed pattern coding and grouped labels into coherent thematic categories. As for the previous extract, "convenience" and "repetition" represented benefits enabled by the technology, so they were developed into a thematic category of "technical affordances." We jointly compared our codes and discussed until consensus was reached. Finally, we agreed on three themes (technical, social/psychological, and educational), each in two dimensions (affordances vs. concerns). In this way, the trustworthiness of the qualitative analysis was ensured by "independent coding, constant comparison of coding, and resolving discrepancies through discussion" (Selvi, 2020, p. 450).

5. Results

5.1 Within-group comparison of pronunciation measures

Table 1 reports the descriptive statistics about the outcome measures by group. To understand whether the groups improved their pronunciation, Wilcoxon Signed Rank tests were run between the pretest and the delayed posttest scores. All three groups improved their comprehensibility (Co-non-ASR: $Z = -4.54$, $p < 0.001$; Co-ASR: $Z = -4.62$, $p < 0.001$; Auto-ASR: $Z = -4.35$, $p < 0.001$), segmental accuracy (Co-non-ASR: $Z = -4.51$, $p < 0.001$; Co-ASR: $Z = -4.38$, $p < 0.001$; Auto-ASR: $Z = -4.01$, $p < 0.001$), and word stress accuracy (Co-non-ASR: $Z = -4.21$, $p < 0.001$; Co-ASR: $Z = -4.22$, $p < 0.001$; Auto-ASR: $Z = -3.44$, $p = 0.001$).

[Insert Table 1 here]

5.2 Between-group comparison of pronunciation measures

Kruskal-Wallis tests showed that the three groups did not differ in the pretest for comprehensibility ($\chi^2 = 0.21$, $p = 0.889$), segmental accuracy ($\chi^2 = 3.81$, $p = 0.149$), or word stress accuracy ($\chi^2 = 3.01$, $p = 0.222$). In the immediate posttest, significant differences were found for comprehensibility ($\chi^2 = 8.85$, $p = 0.012$), segmental accuracy ($\chi^2 = 11.21$, $p = 0.004$), and word stress accuracy ($\chi^2 = 6.44$, $p = 0.04$). Mann-Whitney tests revealed that the Co-ASR group significantly outperformed the Auto-ASR group in comprehensibility ($Z = -2.92$, $p = 0.004$), segmental accuracy ($Z = -3.24$, $p = 0.001$), and word stress accuracy ($Z = -2.33$, $p = 0.02$). The Co-non-ASR group also significantly outperformed the Auto-ASR group in comprehensibility ($Z = -2.10$, $p = 0.036$), segmental accuracy ($Z = -2.04$, $p = 0.042$), and word stress accuracy ($Z = -2.03$, $p = 0.043$). However, the Co-non-ASR and the Co-ASR groups did not differ significantly.

In the delayed posttest, significant differences were again found for comprehensibility ($\chi^2 = 9.91$, $p = 0.007$), segmental accuracy ($\chi^2 = 11.06$, $p = 0.004$), and word stress accuracy ($\chi^2 = 11.08$, $p = 0.004$). Mann-Whitney tests revealed that the Co-ASR group significantly outperformed the Auto-ASR group in comprehensibility ($Z = -2.92$, $p = 0.004$), segmental accuracy ($Z = -3.21$, $p = 0.001$), and word stress accuracy ($Z = -3.35$, $p = 0.001$). The Co-non-ASR group also significantly outperformed the Auto-ASR group in comprehensibility ($Z = -2.43$, $p = 0.015$), segmental accuracy ($Z = -2.35$, $p = 0.019$), and word stress accuracy ($Z = -2.01$, $p = 0.044$). Again, the Co-non-ASR and the Co-ASR groups did not differ significantly.

5.3 Students' perceptions of three mobile-assisted learning conditions

5.3.1 Questionnaire results

Table 2 presents the perception questionnaire results by group. Kruskal-Wallis tests revealed no significant difference among the groups for all the six items. As the mean scores were all above four, it appeared that the students had generally positive perceptions of the mobile-assisted feedback activity. Although the three groups did not differ in their survey responses, these similar perception results might be qualitatively motivated by different reasons (Crowther et al., 2016). As such, we needed to explore the interview data to have a fuller understanding about the students' perceptions about different mobile-assisted feedback sources.

[Insert Table 2 here]

5.3.2 Interview results

Based on the thematic analysis of the interviews, the affordances and concerns of mobile-assisted pronunciation learning could be grouped into three themes: *technical*, *social/psychological*, and *educational*. Table 3 summarizes the findings in each category across three groups. In the following paragraphs, extracts are cited to illustrate the students'

perceptions. For the ease of notating group information, interviewees from the Co-non-ASR group are identified by pseudonyms with an initial N, the Co-ASR group with an initial C, and the Auto-ASR group with an initial A.

[Insert Table 3 here]

Interestingly, although the three groups used different WeChat functions, they all appreciated the same technical affordances: *convenience* and *repetition*. They recognized that the app allowed them to do the pronunciation task in a convenient manner and repeat as many times as they wanted: "It was very convenient. We could talk to each other anytime anywhere…We could practice again and again" (Naomi); "My peer and I conveniently went over the sentences…so many times that the sentences were imprinted on my mind" (Charles); "I repeatedly checked the transcript and found out whether I pronounced the words correctly…It was quite handy" (Anna).

With respect to social/psychological affordances, both the Co-non-ASR and the Co-ASR groups enjoyed the dyadic *interaction* and *rapport* enabled by WeChat. As Nathan stated, "Both of us were quite relaxed. We talked and laughed. I listened to my peer's suggestions and learned how to pronounce the sentences." Similarly, Charles explained that "This was a dyadic chat, so we did not have to face the public scrutiny in a classroom. I wouldn't be embarrassed even though I mispronounced…The discussion was interesting and motivated me to learn the pronunciation."

Furthermore, the dyadic interaction afforded them a *sense of responsibility* that urged them to get committed to the task: "With two students in a group…we can interact with each other and learn from each other…If I do the task on my own, I will be careless and sloppy" (Norah); "Working in a dyad meant having someone to supervise me. I must keep up and stay focused" (Clara). By comparison, the Auto-ASR group tended to develop a *sense of autonomy* through checking app-generated transcripts and consulting resources: "I consulted pronunciation dictionaries if I found that the transcripts were deviated…This was an autonomous learning task…I practiced by myself and paced myself" (Ava).

In terms of educational affordances, all three groups believed that the feedback they received was *immediate*, regardless it was from peers and/or ASR: "When I mispronounced a word, my partner pointed it out straightaway and told me how to correct it…With ASR, I could also instantly see what were correctly pronounced" (Clara); "Based on the transcript, I could immediately see what went wrong. The feedback was specific to particular words" (Abby).

However, the students described two interesting differences about educational affordances. If the students worked with ASR, they tended to pay more attention to *word linking* and *reading speed* because they wanted to find out whether these factors would affect the speech-to-text recognition. As Caroline recounted, "When I linked words in pronunciation, sometimes they were not recognized by the app. I asked my peer to judge whether my linking was correct. She gave me lots of suggestions to appropriately link

words." Similarly, Alex reported that "I needed to control my reading speed. If the speed was too fast or too slow, the words might not be recognized by the app. This developed my skills to master an appropriate reading speed."

If the students worked in dyads, the feedback givers offered *effective scaffolding* tailored to the feedback receivers' proficiency so that the receivers could fully understand how to improve their pronunciation: "When I listened to online dictionaries, the pronunciation was very fast…when I worked with my peer, she slowed down a lot…I found her feedback very clear and helpful" (Naomi); "My peer told me how to pronounce a word syllable by syllable. I then marked up the word accordingly. This was much better than consulting online dictionaries. The pronunciation in the dictionaries was too fast" (Chris).

With respect to technical concerns, both the Co-non-ASR and the Co-ASR groups alluded to *internet connection*, although the issue was deemed minor and temporary: "Sometimes, internet connection was not good…but this was just a little hiccup" (Nancy); "Internet connection was not always stable and our discussion was temporarily interrupted…but overall I felt quite good about the task" (Charlotte). Additionally, both the Co-ASR and the Auto-ASR groups were not sure of the *robustness of ASR*. They might trust their peers or online dictionaries more than ASR. As Chris recounted, "when WeChat could not recognize my pronunciation, I would turn to my peer for ultimate judgment." Similarly, Alex contended that "my pronunciation should be correct because I tried very hard to imitate the pronunciation in the online dictionary. Still, I got one word not recognized by the app."

With regards to social/psychological concerns, both the Co-non-ASR and the Co-ASR groups pointed out the potential issue about their peers' *unwillingness to work* with them: "It depends on whether your partner is willing to cooperate with you. If not, the communication cost is quite high. Luckily, my partner for this task was quite cooperative" (Nicole). By comparison, the Auto-ASR group worried about *not being supervised*: "Without being supervised, I easily got distracted. In the middle of the task, I began to play with my phone and my mind began to wander" (Alice).

Finally, in terms of educational concerns, both the Co-non-ASR and the Co-ASR groups alluded to *peers' proficiency* as a potential issue. As Nicolas stated, "the higher-grade student might not be perfect. It was possible that he did not know how to pronounce a word...What if both of us were wrong?" This was echoed by Cecelia: "Sometimes both of us were not sure of the pronunciation or the word stress. We tried to solve these issues by consulting online dictionaries and practicing with ASR multiple times, but some issues remained unresolved." As for the Auto-ASR group, they wished to receive *more learning supports* to help them identify and address pronunciation issues: "It would be better to have a reference version. We could listen to the sentences and learn from the reference version. This was better than consulting the pronunciation of individual words in dictionaries" (Anthony).

6. Discussion and implications

The within-group comparisons of pronunciation measures show that all three groups improved their pronunciation, albeit to a varying degree. Even within the Auto-ASR group, the students improved their comprehensibility, segmental accuracy, and words stress accuracy. These results are in line with skill acquisition theory (DeKeyser, 2017) in that deliberate practice with mobile-assisted peer and/or ASR feedback contributes to proceduralization of target pronunciation features. These findings not only lend further empirical support to previous research on the effectiveness of ASR in pronunciation training (e.g., Golonka et al., 2014; McCrocklin 2019b),), but also contribute to the line of research on the effectiveness of mobile-based dictation ASR (e.g. Nuance Dragon Dictation in Liakin et al., 2015; Gmail in Mroz, 2018).

Notwithstanding these promising results, the between-group comparisons show that peer feedback was superior to ASR feedback because both the Co-non-ASR and the Co-ASR groups outperformed the Auto-ASR group in the immediate and delayed posttests. Based on the students' interviews, we posit that the discrepancy might be primarily caused by the "effective scaffolding" afforded by the peer feedback (see Table 3 "educational affordances"). As Lantolf (2012) aptly pointed out, for scaffolding to be effective, "the quality and quantity of external forms of social interaction [should be] attuned to a learner's potential ability" (p. 57). The focal participants in the current study were all lower-grade students. For the Auto-ASR group, although dictation ASR offered them immediate and detailed feedback, the feedback was indirect and they had to rely on themselves to figure out how to address the pronunciation issues (e.g. consulting pronunciation dictionaries). On the contrary, the Co-non-ASR and the Co-ASR groups had more capable peers to offer feedback attuned to their ability. For instance, as reported in the interviews, their peers pronounced the words at a much slower speed than the pronunciation dictionaries. Their peers also split the words into syllables and modeled the pronunciation syllable by syllable. These effective scaffoldings could explain why the Co-non-ASR and the Co-ASR groups attained better pronunciation learning.

On a different but related note, it is somewhat surprising to find that the Co-ASR group did not significantly outperform the Co-non-ASR group, even though the Co-ASR group received more input. Two reasons might be possible. First, the higher-grade students in the dyads were able to identify pronunciation issues even without the input from ASR. Hence, *additional* diagnostic information from ASR might not be substantial. As revealed in the interviews, compared to the Co-non-ASR group, the Co-ASR group was more likely to work on word linking and reading speed. This might be triggered by the speech-to-text transcription and the peers tried to figure out whether these elements would skew the transcription. However, this additional diagnostic information was not directly related to the three outcome measures, leading to an impression that the effect of additional ASR feedback was marginal. Second, students' attitudes toward ASR might affect how they used its feedback. As revealed in the interviews, they had some reservations about the technology and tended to have higher regard for peer feedback. As such, even though ASR offered valid feedback, they might not act on it, thus limiting the additional benefits of ASR in the dyadic interaction.

In light of previous findings, our study has three pedagogical implications for mobile-assisted pronunciation learning. First, although the Auto-ASR group was outperformed by the Co-non-ASR and the Co-ASR groups, there is a place for ASR in mobile-assisted pronunciation learning, especially when individual students want to engage in sustained, self-paced, and stress-free practice (Bashori et al., 2020; Neri et al., 2008). It should be reiterated that the Auto-ASR group improved in three outcome measures after the task and that all three groups did not differ in the perception questionnaires. According to skill acquisition theory, practice with feedback has a better effect on proceduralization and automatization than practice alone (Sato & Lyster, 2012). Therefore, the advantages of using dictation ASR for timely and cost-effective feedback should be recognized (Mroz, 2018). To maximize learning potentials, instructors need to make students aware of what the technology can and cannot do. For instance, instructors can point out that dictation ASR in smartphones has more to do with segmental elements rather than suprasegmentals (McCrocklin, 2019a). If students work on a controlled production task (e.g. reading sentences), it is a good idea for instructors to test the speech-to-text recognition and point out which words cannot be recognized. If the recognition is good, we can reassure students of the validity of ASR feedback. This enables them to develop a healthy (rather than an overly negative) attitude towards ASR feedback. In addition, as dictation ASR feedback is indirect and needs to be complemented by other more direct feedback, students should have a relatively high level of autonomy to take full advantage of ASR feedback. Therefore, self-regulated strategies (e.g. noticing and self-monitoring) can be introduced to help students act on ASR feedback. Addressing the concerns raised by our students, two versions of reading samples (one at a normal speed, the other at a slower speed) can be sent to students so that they can listen and model from the samples more attuned to their proficiency. They can practice pronouncing single words until ASR can recognize them, and then practice the whole sentences.

Second, similar to computer-assisted collaboration (Evers & Chen, 2020; Luo, 2016; Tsai, 2019), mobile-assisted collaborative tasks are meaningful sites for students to improve their pronunciation. Given the popularity of social networking apps, peer collaboration is no longer bound by physical proximity. This means that geographically dispersed students can be connected via mobile technologies, making peer collaboration possible even in the midst of a pandemic that requires social distancing (the context of our study). Although some students in our study mentioned that internet connection was occasionally disrupted, their overall experience was good. Working in dyads induced a sense of responsibility and offered important opportunities for the students to accomplish the task with peers' effective scaffolding. Although they did not report communication breakdowns, they alluded to the potential concern of their partners' unwillingness to work with them. This concern can be addressed by instructors, who instill a "collaborative mindset" in students (Sato & Ballinger, 2012) and model interaction strategies. Prior to a mobile-assisted collaborative pronunciation task, instructors can raise students' awareness of the benefits of peer collaboration and appeal to students' interest in repurposing a social networking app (e.g. WeChat and WhatsApp) into a collaborative learning app. Additionally, instructors can demonstrate a variety of interaction strategies, including pre-

task planning, elaborated explanation, joint discovery, and supportive encouragement (Dao, 2020; Tsai, 2019).

Building on the two previous points, we contend that ASR feedback and peer feedback can be combined to complement instructor feedback in what we call a "sandwich model." This echoes Tsai's (2015) proposal to have students engage in technology-mediated self-practice and then in dyadic interaction, which builds into an additional round of self-practice. More specifically, in our "sandwich model," individual students use mobile-based dictation ASR to practice their pronunciation and note down unresolved issues. Then, they use social networking apps (and learning apps) to work on these issues in a collaborative manner. After that, students once again use ASR and autonomously check whether there are remaining issues, which will be addressed by instructors during class time. As such, the "sandwich model" provides an optimal balance of the affordances of autonomy and peer connectivity outside of class (Godwin-Jones, 2011). When the tripartite feedback of ASR, peer, and instructor are carefully orchestrated, it can ease the concerns that students are not supervised (reported by the Auto-ASR group) and that even higher-grade peers may be unable to solve some pronunciation issues (reported by the Co-non-ASR and the Co-ASR groups). In this way, mobile-assisted pronunciation learning, both autonomous and collaborative, can take place outside of class and save in-class time for lingering issues and other skill training activities (Bodnar et al., 2017; McCrocklin, 2019a).

7. Conclusion

This study compared three mobile-assisted learning conditions, in which students received feedback from peer and/or ASR. The findings show that all three groups improved their pronunciation in terms of comprehensibility, segmental accuracy, and word stress accuracy, although mobile-assisted peer feedback seemed to have a better learning effect than ASR feedback. The three groups did not differ in the perception questionnaires. However, their interviews revealed some shared and unique technical, social/psychological and educational affordances and concerns. Based on these findings, we contend that ASR should have a place in mobile-assisted pronunciation learning so that students can receive immediate and individualized feedback in self-paced practice (Bodnar et al., 2017; Pennington & Rogerson-Revell, 2019). We also propose to leverage the increasing ubiquity of social networking apps and the dictation ASR technology embedded in these apps to engage students in autonomous and collaborative pronunciation learning outside of class. In this way, the tripartite feedback of ASR, peer, and instructor can be optimally balanced and sequenced to maximize learning potentials.

One limitation of this study is its primary focus on the controlled production of a finite set of pronunciation features. More research is needed to understand whether and how dictation ASR can also lead to pronunciation improvement in spontaneous speech (e.g. Evers & Chen, 2020). Another limitation is the cross-sectional nature of the study. Previous research has shown that students' attitudes towards learning technology (e.g. Raes & Depaepe, 2020) and collaborative dynamics (e.g. Chen & Yu, 2019) can change over time. Therefore, it is interesting to know how multiple iterations of autonomous and

collaborative pronunciation learning play out and what factors mediate students' longer-term perceptions, which in turn have an impact on their pronunciation learning. Future studies can adopt a longitudinal design and trace students' attitudes towards ASR and collaborative dynamics over multiple pronunciation tasks. These insights will shed important light on the feasibility of mobile-assisted pronunciation learning as a normalized training routine, rather than a peripheral add-on.

References

Bashori, M., van Hout, R., Strik, H., & Cucchiarini, C. (2020). Web-based language learning and speaking anxiety. *Computer Assisted Language Learning*, 1-32.

Bodnar, S., Cucchiarini, C., de Vries, B. P., Strik, H., & van Hout, R. (2017). Learner affect in computerised L2 oral grammar practice with corrective feedback. *Computer Assisted Language Learning*, *30*(3-4), 223-246.

Chang, J. (1987). Chinese speakers. In Michael Swan & Bernard Smith (eds.). *Learner English: A teacher's guide to interference and other problems* (pp. 224-237). Cambridge University Press.

Chen, W., & Yu, S. (2019). A longitudinal case study of changes in students' attitudes, participation, and learning in collaborative writing. *System*, *82*, 83-96.

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.

Creswell, J. W. & Plano Clark, V. L. (2007). *Designing and conducting mixed methods research*. Sage.

Crowther, D., Trofimovich, P., & Isaacs, T. (2016). Linguistic dimensions of second language accent and comprehensibility: Nonnative listeners' perspectives. *Journal of Second Language Pronunciation*, *2*(2), 160-182.

Dao, P. (2020). Effect of interaction strategy instruction on learner engagement in peer interaction. *System*, 102244.

De Vries, B. P., Cucchiarini, C., Bodnar, S., Strik, H., & van Hout, R. (2015). Spoken grammar practice and feedback in an ASR-based CALL system. *Computer Assisted Language Learning*, *28*(6), 550-576.

DeKeyser, R. (2017). Knowledge and skill in ISLA. In S. Loewen & M. Sato (eds.), *The Routledge handbook of instructed second language acquisition* (pp. 15-32). Routledge.

Deterding, D. (2006). The pronunciation of English by speakers from China. *English World-Wide*, *27*(2), 175-198.

Dlaska, A., & Krekeler, C. (2013). The short-term effects of individual corrective feedback on L2 pronunciation. *System*, *41*(1), 25-37.

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.

Godwin-Jones, R. (2011). Emerging technologies: Mobile apps for language learning. *Language Learning & Technology*, *15*(2), 2-11.

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.

Hu, G., & Lam, S. T. E. (2010). Issues of cultural appropriateness and pedagogical efficacy: Exploring peer review in a second language writing class. *Instructional Science*, *38*(4), 371-394.

Huisman, B., Saab, N., van Driel, J., & van den Broek, P. (2017). Peer feedback on college students' writing: exploring the relation between students' ability match, feedback quality and essay performance. *Higher Education Research & Development*, *36*(7), 1433-1447.

Jin, L. (2018). Digital affordances on WeChat: Learning Chinese as a second language. *Computer Assisted Language Learning*, *31*(1-2), 27-52.

Lantolf, J. P. (2012). Sociocultural theory: A dialectical approach to L2 research. In S. M. Gass, & A. Mackey (eds.), *The Routledge handbook of second language acquisition* (pp. 57-72). Routledge.

Levis, J. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly*, *39*(3), 369-377.

Levis, J. (2007). Computer technology in teaching and researching pronunciation. *Annual Review of Applied Linguistics*, 27, 184-202.

Liakin, D., Cardoso, W., & Liakina, N. (2015). Learning L2 Pronunciation with a Mobile Speech Recognizer: French/y/. *CALICO Journal*, *32*(1), 1-25.

Liaw, M. L. (2014). The affordance of speech recognition technology for EFL learning in an elementary school setting. *Innovation in Language Learning and Teaching*, *8*(1), 79-93.

Loewen, S., Crowther, D., Isbell, D. R., Kim, K. M., Maloney, J., Miller, Z. F., & Rawal, H. (2019). Mobile-assisted language learning: A Duolingo case study. *ReCALL*, *31*(3), 293-311.

Luo, B. (2016). Evaluating a computer-assisted pronunciation training (CAPT) technique for efficient classroom instruction. *Computer Assisted Language Learning*, *29*(3), 451-476.

Mackey, A. (2020*). Interaction, feedback and task research in second language learning: Methods and design*. Cambridge University Press.

McAndrews, M. M., & Thomson, R. I. (2017). Establishing an empirical basis for priorities in pronunciation teaching. *Journal of Second Language Pronunciation*, *3*(2), 267-287.

McCrocklin, S. (2019a). ASR-based dictation practice for second language pronunciation improvement. *Journal of Second Language Pronunciation, 5*(1), 98-118.

McCrocklin, S. (2019b). Learners' feedback regarding ASR-based dictation practice for pronunciation learning. *CALICO Journal*, *36*(2), 119-137.

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook (3rd ed.).* Sage.

Mroz, A. (2018). Seeing how people hear you: French learners experiencing intelligibility through automatic speech recognition. *Foreign Language Annals*, *51*(3), 617-637.

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.

Pennington, M. C., & Rogerson-Revell, P. (2019). Using technology for pronunciation teaching, learning, and assessment. In M. C. Pennington & P. Rogerson-Revell (eds.). *English pronunciation teaching and research* (pp. 235-286). Palgrave Macmillan.

Raes, A., & Depaepe, F. (2020). A longitudinal study to understand students' acceptance of technological reform. When experiences exceed expectations. *Education and Information Technologies*, *25*(1), 533-552.

Sato, M., & Ballinger, S. (2012). Raising language awareness in peer interaction: a cross-context, cross-methodology examination. *Language Awareness*, *21*(1-2), 157-179.

Sato, M., & Lyster, R. (2012). Peer interaction and corrective feedback for accuracy and fluency development: Monitoring, practice, and proceduralization. *Studies in Second Language Acquisition*, *34*(4), 591-626.

Saito, K., & Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning*, *69*(3), 652-708.

Saito, Y., & Saito, K. (2017). Differential effects of instruction on the development of second language comprehensibility, word stress, rhythm, and intonation: The case of inexperienced Japanese EFL learners. *Language Teaching Research*, *21*(5), 589-608.

Selvi, A. F. (2020). Qualitative content analysis. In J. McKinley & H. Rose (eds.). *The Routledge handbook of research methods in applied linguistics* (pp. 440-452). Routledge.

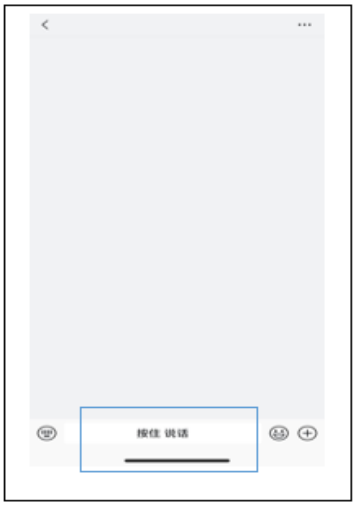
Tsai, P. H. (2015). Computer-assisted pronunciation learning in a collaborative context: A case study in Taiwan. *The Turkish Online Journal of Educational Technology*, *14*(4), 1-13.

Tsai, P. H. (2019). Beyond self-directed computer-assisted pronunciation learning: a qualitative investigation of a collaborative approach. *Computer Assisted Language Learning*, *32*(7), 713-744.

van Doremalen, J., Boves, L., Colpaert, J., Cucchiarini, C., & Strik, H. (2016). Evaluating automatic speech recognition-based language learning systems: A case study. *Computer Assisted Language Learning*, *29*(4), 833-851.

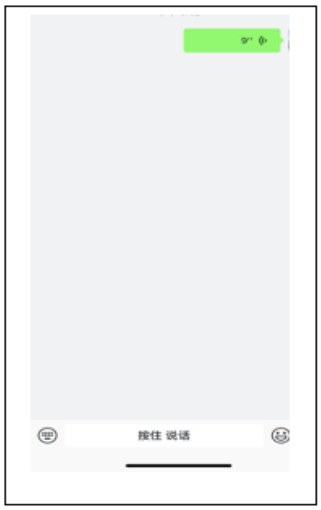
Appendix 1 Description of the dictation ASR in WeChat

Step 1
Press the button (highlighted part) and talk

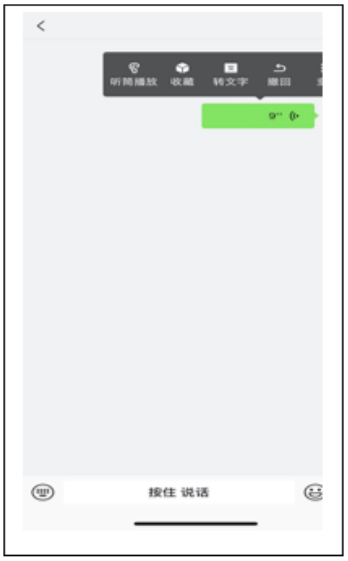


Step 2
When finished, release the button to send the audio

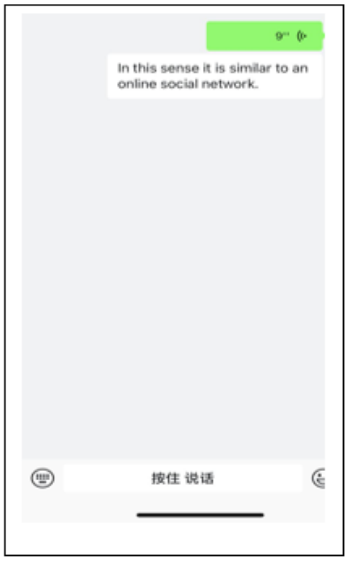
This shows that the audio is sent



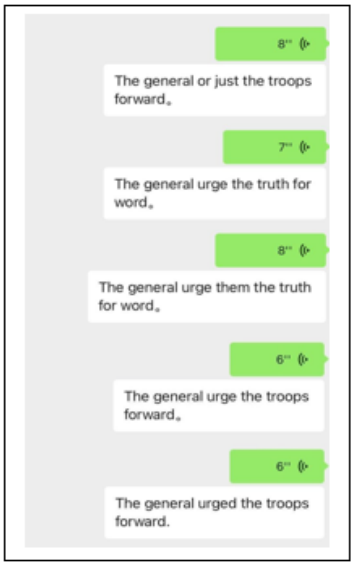
Step 3:
Select the audio and activate the menu bar. Choose "transcribe" (the third option from the left)



This shows the ASR transcript.



Step 4:
Try to practice until the attempt is successful.



Appendix 2 Sentences used for the study

Target vowels and consonants in the assessment of segmental accuracy are provided right next to the words. Bolded words are target stressed words in the assessment of word stress accuracy.

1. In this /ð/ sense, it is similar to an **online** social network /əː/.

2. With /ð/ your knowledge of /v/**agriculture** /ʌ/, you can **transform** /tr/ the land into a vast /v/ /ɑː/ and **productive** /ʌ / /v/ farm /ɑː/.
3. He /iː/ is a **prominent** scholar in the field /iː/ of /v/ **linguistics**.
4. The general urged /əː/ the troops **forward**.
5. Sweat streamed /str/ /iː/ down his face /eɪ/.
6. We /iː/ set out to find the truth /tr/ /θ/ behind the **mystery**.
7. When night fell, the construction /str/ /ʌ/ site was ablaze /eɪ/ with /ð/ light.
8. The /ð/ **actress** /tr/ **dedicated** /eɪ/ her /əː/ life to children's charity work /əː/.

Table 1. Descriptive statistics of the three outcome measures in three tests

Outcome measures	Co-non-ASR		Co-ASR		Auto-ASR	
	Mean	SD	Mean	SD	Mean	SD
Comprehensibility						
Pretest	5.15	1.21	5.11	1.27	5.08	1.04
Posttest	7.35	0.88	7.56	0.67	6.81	1.20
Delayed posttest	7.12	0.90	7.26	0.84	6.53	1.05
Segmental accuracy						
Pretest	20.11	3.84	21.93	4.12	20.36	3.79
Posttest	26.04	2.78	27.04	2.36	24.32	3.41
Delayed posttest	25.32	2.89	26.11	2.22	23.54	3.21
Word stress accuracy						
Pretest	3.18	2.47	3.89	1.81	3.32	2.06
Posttest	6.32	1.47	6.57	1.20	5.54	1.57
Delayed posttest	5.79	1.77	6.39	1.40	4.71	1.90

Table 2. Perception questionnaire results

I think that…	Co-non-ASR		Co-ASR		Auto-ASR	
	Mean	SD	Mean	SD	Mean	SD
the activity has developed my understanding about English pronunciation.	4.32	1.68	4.64	1.13	4.43	1.43
the activity has improved my English pronunciation.	4.32	1.66	4.36	1.22	4.39	1.17
the activity has improved my confidence in English pronunciation.	4.25	1.62	4.14	1.38	4.25	1.24
the activity was interesting.	4.32	1.81	4.39	1.37	4.25	1.27
the activity was meaningful.	4.50	1.62	4.36	1.34	4.36	1.28
I will repeat this activity to practice my English pronunciation.	4.29	1.70	4.21	1.42	4.50	1.23

Table 3. Affordances and concerns reported by the students

Items	Co-non-ASR	Co-ASR	Auto-ASR
Technical affordances			
Repetition	✓	✓	✓
Convenience	✓	✓	✓
Social/psychological affordances			
Interaction & rapport	✓	✓	
Sense of responsibility	✓	✓	
Sense of autonomy			✓
Educational affordances			
Immediate feedback	✓	✓	✓
Effective scaffolding	✓	✓	
Attention to word linking and reading speed		✓	✓
Technical concerns			
Internet connection	✓	✓	
Robustness of ASR		✓	✓
Social/psychological concerns			
Unwillingness to work with peers	✓	✓	
Not being supervised			✓
Educational concerns			
Peers' proficiency	✓	✓	
More learning support needed			✓