СОВРЕМЕНА ФИЛОЛОГИЈА
УДК 004.934.2:811.111’243

JOURNAL OF CONTEMPORARY PHILOLOGY
JCP 2020, 3 (2), 45-61.

ASR-DICTATION ON SMARTPHONES FOR VOWEL PRONUNCIATION PRACTICE

Agata Guskaroska
Iowa State University
agatag@iastate.edu

This study aims to explore mobile-assisted Automated Speech Recognition (ASR) dictation systems for vowel pronunciation practice by examining whether ARS can be useful for pronunciation improvement and speech recognition accuracy. Additionally, learners’ attitudes towards using these systems were explored. Twenty-one Macedonian EFL learners practiced pronouncing 26 words with the following minimal pairs: /i/, /ɪ/; /æ/, /ɛ/; /u/, /ʊ/; /ɑ/, /ʌ/. The participants were divided into an experimental group (n=11) and a control group (n=10). This study used a mixed methods approach including qualitative and quantitative analysis. Results demonstrated that while the control group did not show any improvement, the experimental group improved their accuracy. ASR written output and human judgment was also found to be within an acceptable agreement for most vowels. Furthermore, while occasional inaccurate feedback sometimes caused frustration, ASR training was generally enjoyed and considered as a practical and safe environment for practice. The findings provide some support for the use of ASR in EFL classrooms with careful planning and direction from the teacher. Using ASR as a tool for controlled and structured practice with individual words is particularly applicable when the focus is to raise learners’ phonological awareness and perception of English vowel sounds.

Keywords: ASR, CALL, MALL, pronunciation practice, EFL

ОБУКА СО ПРОГРАМАТА ЗА АВТОМАТСКО ПРЕПОЗНАВАЊЕ ГОВОР ЗА ПОДОБРУВАЊЕ НА ИЗГОВОРОТ НА ВОКАЛИТЕ

Агата Гускароска
Државен универзитет во Ајова
agatag@iastate.edu

Студијата има за цел да ги истражи системите за автоматско препознавање говор и нивната улога во подобрување на изговорот на вокали и нивото на препознавање на изговорот. Исто така, оваа студија истражува како учесниците гледаат на овие системи. Имено, 21 изучувачи на англискиот јазик учествуваа во оваа студија. Македонските изучувачи на англискиот јазик го вежбаа нивниот англиски изговор преку изговор на 26 зборови со следните минимални парови: /i/, /ɪ/; /æ/, /ɛ/; /u/, /ʊ/; /ɑ/, /ʌ/. Учесниците беа поделени во експериментална (11) и контролна група (10). Оваа студија користеше квалитативна и квантитативна анализа на податоците. Резултатите покажаа дека експерименталната група покажа подобрување на изговорот додека контролната група не покажа напредок. Беше утврдено дека излезниот напишан текст (autput) од програмата за автоматско препознавање на говор беше во рамките на прифатливо отстапување споредено со напишаниот текст од изучувачите. Исто така, иако изучувачите искажаа повремена фрустрација со погрешни резултати од страна на програмата, во главно, учесниците искажаа задоволство од обуката и изјавија дека оваа обука обезбеди практично и безбедно место за вежбање на нивниот изговор. Овие резултати даваат поддршка за употребата на програмата за автоматско препознавање на говор во наставата на англиски јазик. Истото е возможно со внимателно планирање и дирекција од страна на наставникот. Користењето на оваа програма како алатка за контролирана и структурирана обука со индивидуални зборови е особено применлива за подигање на фонолошката свест и перцепција на англиските вокали.

Клучни зборови: автоматско препознавање говор, изучување јазик со помош на компјутерска технологија, изучување јазик со помош на мобилни уреди, вежбање изговор, англиски како странски јазик

1 Introduction

Pronunciation is often a neglected area in the second language (L2) classroom due to lack of time, difficulties in providing individual feedback (McCrocklin 2016), lack of resources, lack of confidence in teaching pronunciation, and uncertainty about how to integrate pronunciation into the curriculum (Levis and Grant 2003). Small countries, such as Macedonia, present an EFL setting where exposure to native speech is not common. Large classes, lack of teacher preparation, and limited resources, very often lead to neglecting pronunciation in regular EFL classes. Therefore, investigating practical tools that can be used in a classroom setting is highly needed in these types of settings. While there are many available digital tools, feedback is what is mostly lacking in computer-assisted pronunciation training (CAPT). Automated Speech Recognition (ASR) is an intriguing software that may offer possibilities for pronunciation practice in an L2. Levis and Suvorov (2013) define ASR as "an independent, machine-based process of decoding and transcribing oral speech" (2014: 1) which turns the speech signal into text. While ASR can be used for several purposes, its use for language learning has considerably increased throughout the years (Ahn and Lee 2016). The potential of ASR to identify pronunciation problems and give feedback to learners has been of interest to many researchers (Cucchiarini et al. 2000; Eskenazi 1999). While earlier studies were not in favor of ASR (Coniam 1999; Derwing et al. 2000), more recent studies show support for the use of ASR for pronunciation practice (McCrocklin 2019; Mroz 2018). Researchers have investigated ASR-based CAPT systems (for example, Dutch CAPT) and commercial ASR dictation systems (for example, Siri). While ASR-based systems with explicit feedback can be important, the ASR-dictation systems are easily and freely available on smartphones and may find a broader use by EFL teachers around the globe.

Even though a few studies have explored pronunciation practice using ASR-dictation systems on smartphones (Liakin et al. 2014; Mroz 2018), the topic needs a lot more exploration. Therefore, this study aims to investigate the potential of mobile-assisted ASR dictation systems for pronunciation improvement. The purpose of the study is to explore: 1) the accuracy of ASR, 2) vowel production improvement after using ASR for pronunciation practice, and 3) learners' attitudes towards ASR.

2 Literature review

2.1 ASR accuracy for providing feedback

An ideal ASR system will identify learners' errors at the same level as humans' perception. However, the holy grail of a computer that matches human speech recognition is still out of reach (Levis and Suvorov, 2013). Methods used to evaluate ASR accuracy typically involve a comparison between ASR written output of native speakers (NS) and non-native speakers (NNS) and/or a comparison between

ASR written output of NNS and native speakers' ratings of the same speech. ASR written output is considered successful if it is similar to native speakers' judgments.

Early research found lower recognition for NNS speech and concluded that ASR was not accurate enough for pronunciation practice (Coniam 1999; Derwing et al. 2000). Nonetheless, improvements in ASR technology have led to continuing exploration of the potential of these systems and their application in L2 classrooms. While many studies have found up to 95% recognition for NS speech (Ehsani and Knodt 1998), the systems' recognition of NNS speech still remains much lower. Even though limitations exist, a few studies show a close relationship between ASR and human judgments (Cucchiarini et al. 2009; Neri et al. 2002). ASR may not always be 100% accurate, but studies show that learners may still improve.

2.2 ASR and pronunciation improvement

Speaking practice offers possibilities for learners to notice gaps in their L2 pronunciation. Swain (1985) suggested that language production was important in three main ways: (a) triggering noticing when learners realize their speech does not effectively send the intended message; (b) testing the speech production when learners try out a way of saying something and then receive feedback; and (c) providing a basis for metalinguistic reflection when learners consciously think about what they have said. Feedback provided by ASR's written output may help improve learners' pronunciation by providing the correct language form, and hence, trigger noticing, conscious thinking and raising students' awareness of the differences between L2 speech sounds that are difficult to perceive and produce.

Exploration of pronunciation improvement with ASR training mostly showed positive results. A few studies, (such as Cucchiarini et al. 2009; Neri et al. 2008) investigated ASR-based CAPT systems by making a comparison of learners' speech before and after an ASR training period, while other studies explored commercial ASR dictation systems (e.g. Liakin et al. 2014; Mroz 2018). Cucchiarini et al. (2009) explored the ASR system Dutch-CAPT and found that the experimental group using ASR improved the most. Similar results were found by Liakin et al. (2014) exploring French L2 learners' pronunciation of the sound /y/ using a commercial ASR application (Nuance Dragon Dictation). Mroz (2018) also looked at French L2 learners and using ASR to raise students' awareness of their intelligibility. Her findings indicated that the learners improved their pronunciation by using ASR dictation system Gmail and the French language pack. Given that ASR dictation systems can be used on mobile phones, their accessibility, familiarity, and practicability can offer potential advantages for L2 learning such as a self-paced learning approach (Victori and Lockhart 1995). The use of mobile-assisted ASR in the L2 classroom seems to be beneficial and practical for pronunciation practice, but it still needs further investigation.

2.3 Learners' benefits and attitudes towards ASR

Research has shown that ASR for pronunciation practice brings numerous learners' benefits (McCrocklin 2016; Mroz 2018). Raising learners' awareness of their speech is a valuable step that can allow learners to monitor and correct their own errors (Ahn and Lee 2016). Being intelligible is important for successful communication and ASR can serve as a way to discover "how people hear you" (Mroz 2018: 1). Most importantly, learners appreciate the use of ASR because they can produce more output in a low-anxiety environment (Chen 2011). Previous studies reported that, in general, the learners' attitudes towards ASR systems were positive (Ahn and Lee 2016; Chen 2011). The learners generally believe in the usefulness of ASR-facilitated training (Cucchiarini et al. 2009). Occasional students' frustration when receiving incorrect feedback was also noted (McCrocklin 2014) but despite these drawbacks, the overall results show that ASR creates a safe space for learners, allowing them repeated practice (McCrocklin 2016). Past literature has shown that students mostly enjoy ASR training, but these issues need further exploration to confirm and strengthen these findings.

2.4 The Study

Previous studies on ASR for pronunciation learning have mostly focused on evaluating ASR CAPT systems while very few studies explored ASR dictation systems on mobile devices. A growing number of students own smartphones that have ASR dictation program installed in the system. Inspired by the lack of research, the increasing use of smartphones worldwide, as well as their ubiquity, the current study explores mobile-assisted dictation ASR programs. Inspired by the lack of research and the increasing use of smartphones worldwide, the current study explores the applicability of mobile-assisted dictation ASR programs for pronunciation practice. More specifically, it investigates its effectiveness by combining evaluations of ASR accuracy and pronunciation improvement as well as learners' attitudes towards ASR. Hence, the study aims to explore the following research questions (RQ):

RQ1: How accurate ASR dictation systems are compared to human raters?

RQ2: To what extent do learners improve their pronunciation of vowels after using ASR?

RQ3: What are the learners' attitudes towards using ASR for pronunciation training?

3 Methods

A mixed method approach was used – quantitative data to answer RQs1 and RQ2 and qualitative data to answer RQ3. Figure 1 summarizes the methods and types of analysis used to test the usefulness of ASR for pronunciation practice.

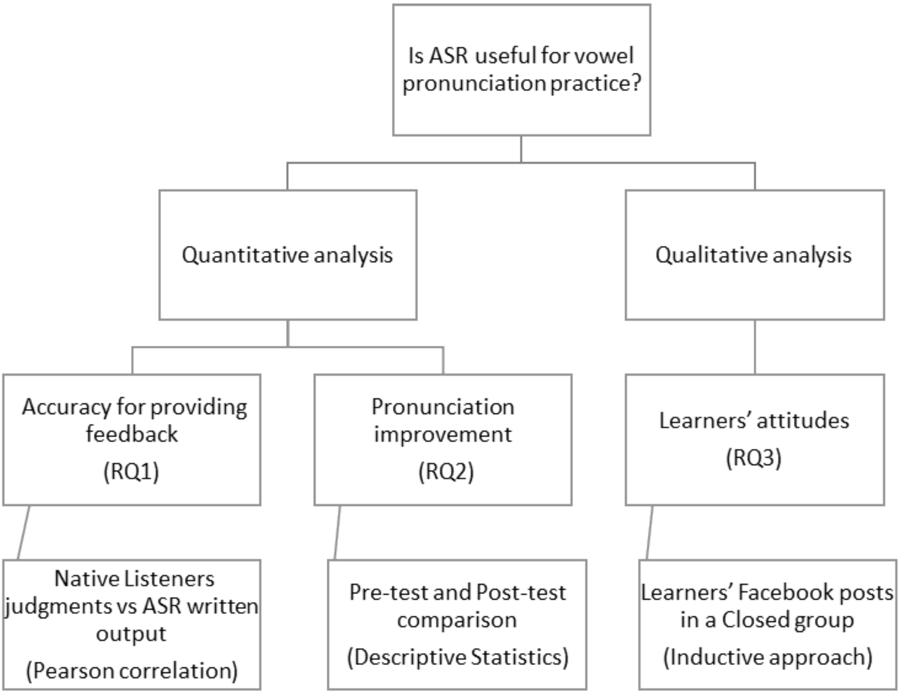
Is ASR useful for vowel pronunciation practice?
Quantitative analysis
Qualitative analysis
Accuracy for providing feedback (RQ1)
Pronunciation improvement (RQ2)
Learners' attitudes (RQ3)
Native Listeners judgments vs ASR written output (Pearson correlation)
Pre-test and Post-test comparison (Descriptive Statistics)
Learners' Facebook posts in a Closed group (Inductive approach)

Figure 1. Summary of Research Methods

3.1 Participants

To answer RQ1, two groups of participants were involved: native listeners and non-native speakers that belonged to the experimental group (EXP). The listeners who transcribed the non-native speech were 10 American native speakers (M=4, F=6). They were all graduate students in an applied linguistics department and had previously taken a pronunciation-related course. The EXP group included 11 Macedonian learners of English, (M=8, F=3, mean age = 18.6), studying English for an average of 10.3 years. They completed a linguistic background questionnaire stating that none of them had visited an English-speaking country and had no prior pronunciation training, but they had continuously been studying English since they were six. Based on self-reported data, there were five intermediate and six high level students.

To answer RQ2, the participants were divided into two groups, the experimental group (EXP) and a control group consisting of non-native speakers (NNSC). NNSC group included 10 Macedonian learners of English (M=6, F=4, mean age = 19.6). Same as EXP, the participants had no prior pronunciation training. They were all learners of English, enrolled in an English course. Their self-reported level was three intermediate and seven high level students. Finally, to answer RQ3, qualitative data was obtained from the participants that belong to the EXP group, described in detail above.

3.2 Stimuli

This study compared the Macedonian and American English vowel systems referring to Flege's Speech Learning Model (SLM) (Flege 1995; 2007) and relevant literature about Macedonian EFL learners' issues (Kirkova-Naskova 2010; 2012) to hypothesize which sounds should be included in the training. Flege's SLM hypothesizes that the closer an L2 sound is to an L1 sound category, the more difficult it will be for the learners to establish a new category for it. Accordingly, learners can assimilate some sounds to an already existing category in the L1. As illustrated in Figure 2, the phonetic system of Macedonian includes five vowels: /i/, /e/, /a/, /o/ and /u/, whereas in English there are arguably around 12 vowels and eight diphthongs (Dodd and Mills 1996)

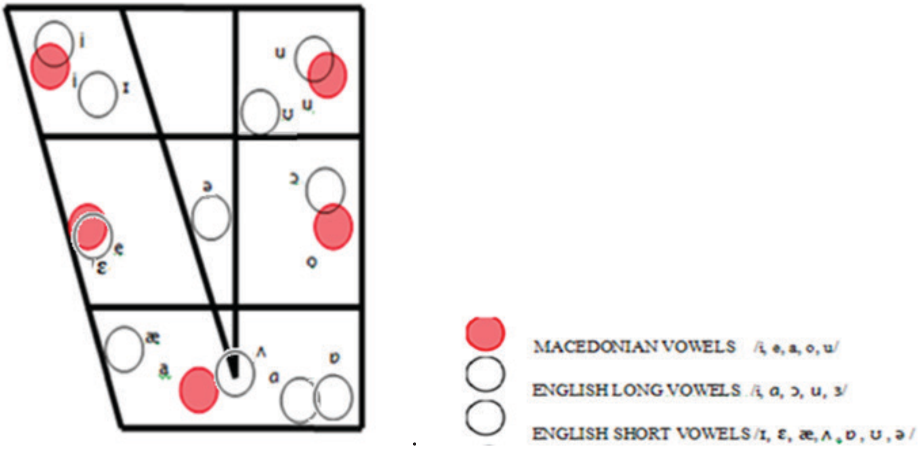
i
i
ɪ
u
u
ʊ
ə
ɔ
e
ɛ
o
æ
a
ʌ
ɑ
ɒ
MACEDONIAN VOWELS /i, e, a, o, u/
ENGLISH LONG VOWELS /i, ɑ, ɔ, u, ɜ/
ENGLISH SHORT VOWELS /ɪ, ɛ, æ, ʌ, ɒ, ʊ, ə /

Figure 2. Macedonian and English vowel diagrams (adapted from Krikova-Naskova 2012)

This study investigates American vowels.[1] Almost every English vowel presents a potential pronunciation problem for Macedonian learners (Kirkova-Naskova 2012). It is very likely that learners might assimilate the English long and short vowels into one category. For example, the English /i/ and /ɪ/ are often categorized as the Macedonian /i/ which falls somewhere between these two sounds and is acoustically similar but qualitatively different i.e. it has different articulatory features. Based on the comparison between these two systems, this study will analyze the following vowel contrasts: /i/-/ɪ/; /æ/-/ɛ/; /u/-/ʊ/; and /ɑ/-/ʌ/. During the ASR practice, this study focuses on improving pronunciation of these selected individual vowels using minimal pairs.

[1] https://easypronunciation.com/en/american-english-pronunciation-ipa-chart.

3.3 Task

The task consisted of practicing pronunciation of the selected vowels, by using a given vocabulary list in ASR (GBoard[2]; Siri or voice search on smartphones). The initial purpose was to explore only GBoard, however, the Institutional Review Board (IRB)[3] suggested providing freedom for students to choose their preferred ASR tool to protect their privacy. Nonetheless, the students were required to use mobile-assisted tool that is free. The EXP members were instructed to practice 20 minutes a day over two weeks, by pronouncing the given words and referring to the written output of the program as feedback for their pronunciation. The vocabulary list consisted of the following words:

/ɪ/, /i/ - *live/ leave; fill/ feel; ship/ sheep*;
/ɛ/, /æ/ - *pen/pan; left/laughed; bed/ bad*;
/ʌ/, /ɑ/ - *cup/cop; duck/dock; shut/shot*
/u/, /ʊ/ - *full/fool; pull/ pool; look/Luke*.

The vocabulary list contained all the target vowels and six distractors for the initial recordings. Individual words were used instead of sentences or phrases in order to possibly avoid the program's 'assumption out of context'. The location for task performance was not specified, but participants were advised to practice in a quiet place. The participants were informed and explained that their participation in this study would include two recordings, one before and after they have practiced pronunciation using an ASR dictation program on their smartphones for a period of two weeks. A Facebook group was created with all EXP participants. Even though, as IRB recommended, this part was optional assuming that some may not have had a Facebook account, or may not have wanted to participate in an online discussion, everyone decided to do so.

3.4 Procedure

The research procedure for data collection included four phases:

(1) An online questionnaire followed by pre-test (recordings of EXP and NNSC speech);

(2) A treatment period for EXP (practice using ASR) and Facebook group posts;

(3) Post-test (second set of recorded speech samples); and EXP provided ASR written output

(4) Listeners transcribed the recordings.

In Phase 1, EXP and NNSC provided linguistic background information by answering a questionnaire via Qualtrics[4]. Then, they completed the pre-test phase

[2] GBoard is a virtual keboard app developed by Google for Android and iOS devices that has a speech-to-text recognition system.

[3] IRB is an established committee that reviews and approves all research involving human participants in the USA.

[4] For additional information, click here

(recording speech samples of the vocabulary reading a list of words, recorded with iPhone microphone, using normal pace and quiet background). EXP was given instructions how to use ASR for practicing. NNSC was not given any pronunciation instructions. Phase 2 was the period of practice for EXP. The students were explained how to use ASR for pronunciation practice using GBoard, Siri or voice search feature on their smartphones. The students practiced their pronunciation of words containing the vowel minimal pairs, for a period of two weeks, 20 minutes a day. The students were also informed they should focus on producing the correct vowel. The written output of ASR served as an indicator of their mispronunciations. If ASR transcribed the word incorrectly, that was considered as feedback for mispronunciation to the learners. The participants practiced individually, at their convenient time and place and occasionally posted updates about their progress and experience in the Facebook private group. Phase 3 was similar to Phase 1. Both EXP and NNSC completed the post-test, that is, recording speech samples of the same list of words containing vowel minimal pairs (same as the pre-test). The learners from EXP also provided their ASR written output, that is, the words which ASR displayed when they were talking. Finally, in Phase 4, the native listeners transcribed all the recordings from the pre-test and the post-test. They were given instructions to write down the words as they hear them, in normal English spelling.

3.5 Data Analysis

For RQ1, ASR accuracy for transcribing the correct vowel sounds was measured by making a comparison between the ASR written output of EXP and the native listeners' transcription of their speech. A difference of 10-12% was set as an acceptable difference. Descriptive statistics was used to calculate the correctly identified vowels sounds. Finally, Pearson correlation was used to measure the relationship between ASR and human judgment. In order to evaluate ASR usefulness for learners' pronunciation improvement, the recorded speech samples of both EXP and NNSC were transcribed by native listeners in both pre-test and post-test. Same as above, the accurately transcribed vowels were calculated using descriptive statistics. The focus of the analysis was on the vowel sounds, each sound used in three words (/ɛ/ in 'bed', 'left' and 'pen') per speaker; summing up to 33 instances total per one vowel sound for EXP, and 30 instances for NNSC, each transcribed by 10 native listeners. The total number of tokens analyzed was 2640 (EXP) and 2400 (NSCG). The learners' issues with consonants sounds were disregarded. For example, if the listener transcribed 'leaf' and the target word was 'leave' the instance was considered as correct because the vowel sound was correctly identified. Finally, the overall accuracy was calculated for the pre-test and post-test. The results from the pre-test and the post-test were measured and then compared to establish whether the learners had improved after the practice period.

To answer RQ3, the last part of the analysis included qualitative analysis of students' Facebook posts and comments about their experience and attitudes towards ASR's usefulness. A general inductive approach was used (Thomas 2003) in which the raw data emerged from important key themes. The students' posts were coded for emerging themes, concepts and beliefs, and hence conclusions were drawn.

4 Results

4.1 ASR Accuracy for vowel pronunciation practice

To answer RQ1 of how accurately the ASR relates to human raters, the learners' written output in the ASR program was compared to human judgments by comparing the mean recognition scores for all vowels and per vowel.

Table 1. Mean Recognition Scores

	ASR Recognition Scores		Native Listeners' Transcription	
Speakers' L1	M	SD	M	SD
Macedonian (N=11)	56.75	18.34	65.04	23.36

Findings show that the overall ASR recognition of Macedonian speech was 56.75% while the native listeners' transcription recognized 65.04% (Table 1) with a difference of 8.29%. Besides the overall scores, this study explored individual vowel recognition (Table 2).

Table 2. ASR and Native listeners' recognition scores

Lexical items	Leave Feel Sheep	Live Fill Ship	Pan Laughed Bad	Pen Left Bed	Luke Pool Fool	Look Pull Full	Cop Dock Shot	Cup Duck Shut
Vowels	i	ɪ	æ	ɛ	U	ʊ	ɑ	ʌ
ASR M (SD) (n=264 tokens)	75.76 (2.61)	45.45 (2.97)	33.33 (3.17)	78.79 (2.34)	66.67 (2.58)	36.36 (2.27)	42.42 (2.61)	75.76 (2.27)
Native Listeners M (SD) (n=2640 tokens)	74.24 (15.99)	55.76 (15.28)	29.09 (15.35)	94.24 (4.97)	71.52 (18.39)	66.67 (15.63)	32.73 (21.67)	96.06 (9.87)

Findings showed that ASR recognition was closer to native listeners for some of the vowels, while it showed lower recognition for others. The vowels /i/, /æ/, /u/ were similarly recognized from both with a difference of 1.52%, 4.24% and 4.85%, respectively. The difference for /ɪ/ was 10.31% and /ɑ/ it was 9.69%, which also falls into the acceptable difference defined in this study (10-12% difference). However, listeners had a higher recognition than ASR for /ʊ/ (30.31%) and /ɛ/ (15.45%). The situation was similar with /ʌ/, where native listeners recognized 96.06%, while ASR recognized only 75.76%. The number of tokens analyzed by native listeners is higher because the speech samples were rated by 10 native listeners. Next, a Pearson correlation was computed to assess the relationship between these two variables. The Pearson correlation (r), r=0.84 showed that these two variables are closely related. Even though the native listeners showed greater recognition, the ratings followed a similar trend (Figure 3) which suggests that ASR's non-recognition may be identifying actual mispronunciations.

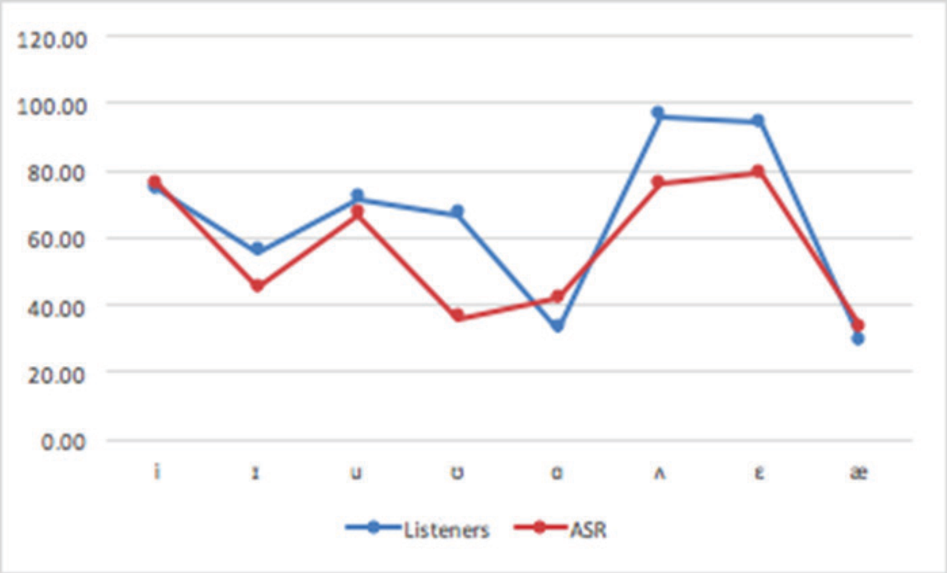

Figure 3. ASR and Native listeners' recognition scores

4.2 Vowel pronunciation improvement

To answer RQ2, results from the pre-test and post-test (Table 3) showed that while EXP improved their overall accuracy score, the NNSC did not show any improvement.

Table 3. Overall accuracy scores on pre-test and post-test (M and SD)

	Pre-test		Post-test	
	M	SD	M	SD
NNSC	62.08	22.77	61.92	25.85
EXP	58.33	25.33	65.04	23.36

Learners who used ASR (EXP) showed different results from the learners who did not have any pronunciation training (NNSC). While NNSC showed only 0.16% difference between pre-test and post-test, the EXP improved 6.71% difference which suggests that the training with ASR may have resulted with certain progress. While the overall (global) improvement is important, investigation of individual vowels can show in-depth analysis (Tables 4 and 5).

Table 4. Learner's pronunciation improvement per individual vowels (NNSC)

Lexical items	Leave Feel Sheep	Live Fill Ship	Pan Laughed Bad	Pen Left Bed	Luke Pool Fool	Look Pull Full	Cop Dock Shot	Cup Duck Shut
Vowels	i	ɪ	Æ	ɛ	u	ʊ	ɑ	ʌ
Pre-test M (SD) (n=2400 tokens)	77.67 (8.55)	49.00 (11.50)	31.67 (13.09)	86.33 (10.96)	60.33 (18.36)	58.00 (24.77)	34.00 (13.26)	99.67 (0.91)
Post-test M (SD) (n=2400 tokens)	79.33 (11.88)	50.67 (15.85)	26.00 (11.64)	95.00 (6.94)	56.33 (18.55)	62.00 (20.61)	27.67 (15.51)	98.33 (2.03)

The results from the NNSC (Table 4) show that there is some difference between the pre-test and the post-test but the differences among individual vowels are inconsistent. More specifically, in a few cases, such as /æ/, /u/, /ɑ/ and /ʌ/ the post-test results are slightly lower than the pre-test. As for the sounds /i/, /ɪ/ and /ʊ/ the participants show some improvement (1.66%; 1.67% and 4%), while a bit higher improvement (8.67%) was noted regarding the sound /ɛ/. Overall, the NNSC's individual scores point out inconsistent increase, decrease or show no improvement in the score.

Table 5. Learner's pronunciation improvement per individual vowels (EXP)

Lexical items	Leave Feel Sheep	Live Fill Ship	Pan Laughed Bad	Pen Left Bed	Luke Pool Fool	Look Pull Full	Cop Dock Shot	Cup Duck Shut
Vowels	i	ɪ	Æ	ɛ	u	ʊ	ɑ	ʌ
Pre-test % M (SD) (n=2640 tokens)	73.64 (16.09)	49.09 (18.17)	16.36 (9.88)	93.94 (5.42)	58.18 (17.50)	64.55 (15.22)	26.06 (18.72)	84.85 (13.82)
Post-test % M (SD) (n=2640 tokens)	74.24 (15.99)	55.76 (15.28)	29.09 (15.35)	94.24 (4.97)	71.52 (18.39)	66.67 (15.63)	32.73 (21.67)	96.06 (9.87)

Findings of EXP demonstrate that the learners improved their pronunciation of each vowel (Table 5). Even though certain improvements were higher than others, it's important to note that there is improvement for each vowel. With a difference of 13.34%, /u/ appears to be the sound where the participants improved the most, followed by /æ/ with 12.73%, and /ʌ/ with 11.15%. Other improvements were found in the sounds /ɪ/ and /ɑ/, both with 6.67% improvement. Finally, the improvement of the sounds /i/ and /ɛ/ were 0.6% and 0.3%, respectively. Even though the practice period was short, certain improvements are clearly visible from the comparison of the pre-test and post-test results, especially regarding the sounds /u/, /æ/, and /ʌ/. The notable improvement in the sound /u/ is important, because the English sound /u/ partially overlaps with the Macedonian /u/, indicating that it could cause difficulties in creating a new category for this sound for the learners. The learners showed the lowest scores for /æ/ which appears to be the most problematic sound for Macedonian EFL learners. The learners were the most accurate with /ε/, (93.94% pre-test and 94.24% post-test) which is almost completely overlapping with the Macedonian /e/; and the sound /ʌ/, whose pronunciation even further improved (11.15%) after the practice period.

4.3 Learners attitudes and beliefs towards ASR

To address RQ3, EXP discussed their experiences in a small closed Facebook group. Results showed that the participants' attitude towards vowel pronunciation practice with ASR was generally positive. At the beginning, most of the learners were initially skeptical about the use of the program because they thought ASR is not highly reliable. Nonetheless, after receiving the instructions and explanations on how to use ASR with the goal to practice their vowels, most of the students reported that they knew 'what to focus on'. Their comments, in the Facebook group showed satisfaction when they became successful with 'getting the program to write the word' they were practicing. One of the learners said: "I kept trying to get the program to write the word pan. I said it more than 20 times… and yes I succeeded and I was so proud of myself. I realized that I was pronouncing the words pen and pan the same. I didn't even know there is a difference between them." Some of the students reported that they became aware of the existence of minimal pairs, and several of them reported that they looked up the words on YouTube to 'hear the difference.' One learner commented: "… now at least I know they are different! I looked up the words on Youtube and online dictionaries…" Another relevant theme that emerged was that the participants appreciated having ASR available at any time and being able to privately practice, without "being judged". Receiving pronunciation feedback in front of their peers would make them feel demotivated to use English in class, and therefore they found using ASR to be their preferred method for receiving feedback. They also agreed that this tool is practical because they all have 'some kind of ASR' already installed on their phones. A few of them reported that they started using ASR as a means of communication with their friends, instead of texting (typing) because they thought it was 'fun.'

Even though the general attitude was positive, occasional frustration was also reported. Four students reported that when the program would write a completely different word which made them feel unsure whether they were mispronouncing the word or the program had problems recognizing it. Two other students commented that they thought the program was more useful when the words were used in sentences. For example, one of them said: "I couldn't write the word pool right, it was either pull or oh or call or nothing on the screen. When I said a sentence like *I jumped in the pool* then it was fine."

5 Discussion

The results of this study showed that ASR has acceptable level of accuracy; learners showed improvements in their pronunciation of vowels after a training period; and the learners showed generally positive attitudes towards its use. Given that ASR and human judgments had similar recognition (8.29% difference), suggest that human raters marked similar mispronunciation errors as ASR. In other words, it is likely that ASR's inaccuracy of transcribing the non-native speech might have been pointing out the learners' mispronunciation because it was similar to the human judgment of inaccurate vowels. Overall scores are a suitable measure to assess ASR accuracy because the ultimate goal is to improve overall pronunciation quality, but, it is also important to provide more in-depth analysis when assessing ASR systems (Neri et al. 2006). The finding that ASR showed an acceptable level of recognition for most of the individual vowels was supported by Coniam (1999) who found that ASR goes consistently in line with human raters and it is most reliable on a word level. This suggests that ASR recognition scores might be a good indicator of the learners' mispronunciations, as they aligned closely with the native listeners' judgments for most of the vowels.

Swain's Output Hypothesis (Swain 1985) suggested that language production was important in three main ways: (1) triggering noticing, (2) receiving feedback, and (3) conscious thinking about the speech production. The language production in the current study may have led to noticing of different vowel contrasts and conscious thinking about their pronunciation of L2 vowels. Pronunciation improvement was also pointed out by Mroz (2018) who found that the learners improved their pronunciation after using a mobile-assisted tool. Other studies that support similar findings regarding segmental improvement are Cucciarini et al. (2009) and Neri et al. (2008). While other studies mostly compared the overall improvement of the learners' speech, this study looked at the improvement of each of the eight selected vowels /i/, /ɪ/; /æ/, /ɛ/; /u/, /ʊ/; and /ɑ/, /ʌ/. Looking at each individual vowel can provide a better insight into the learners' actual improvement. Liakin et al. (2014) explored the French sound /y/ and found that the learners improved after using mobile-assisted ASR. Nonetheless, while longer-term improvement would depend on many other factors, including students' aptitude, attitude, personal motivation and willingness to improve, it is crucial to start by raising the awareness of the existence of unfamiliar sounds. The learners' attitudes are also an important factor in evaluating the usefulness of ASR. The learners' initial skepticism in using ASR was likely due to their expectations of lower recognition of non-native speech. Nonetheless, the 'flaw' of ASR with low non-native speech recognition was used as a feedback of mispronunciation in this study. Learners reported searching for outside resources in an attempt to find native listeners' input which may be an indicator that ASR promoted more autonomous learning, also supported by McCrocklin (2016). The self-reported awareness raising about vowel pronunciation might have allowed them to observe and correct their own errors (Ahn and Lee 2016) and to look for additional resources to improve their perception of the words.

ASR can be useful for both ESL and EFL context. Nonetheless, ASR can be especially useful for EFL learners where exposure to native language is limited (McCrocklin 2016). Even though occasional frustration was mentioned, the students reported more benefits than drawbacks from its use. It is very likely that having an option to practice in the privacy of their home contributed to adopting a more positive attitude toward the use of ASR. Due to lack of exposure to native speech and lack of corrective feedback, ASR may be a useful tool to raise students' awareness of mispronunciation of vowels.

6 Conclusion

This study explored the use of mobile-assisted ASR by investigating three aspects of its usefulness: the accuracy of the program as compared to human judgments, the degree to which ASR can promote improvement, and the learners' attitudes towards using the program for vowel pronunciation practice. The findings indicate that the EXP improved their global pronunciation as well as most of the individual vowels. The findings also showed that ASR's written output was closely related to human judgment with an acceptable difference between them. Nonetheless, the comparison of human judgment and ASR recognition was not uniformly similar for all vowels. Despite these limitations, the students appeared to have positive attitudes towards ASR and enjoyed the training. ASR-dictation systems on smartphones appear to be a practical and useful approach for pronunciation practice. This study recommends the use of ASR-dictation systems in EFL context to address the lack of corrective feedback in this area.

6.1 Implications, limitations and future research

While ASR-dictation programs still have room for improvement, this type of training can serve as an awareness-raising tool that could be integrated into the L2 classrooms. Therefore, this study recommends careful guidance from the teachers, focused and structured practice using individual words. Nonetheless, ASR dictation practice should not be a complete substitute for classroom instruction (McCrocklin 2019) but it can serve as a tool to free up classroom time and facilitate more autonomous practice at home. The teachers need to provide clear goals and guidance to make this type of practice work. For future studies, exploring only one type of ASR would give a clearer picture of the accuracy of that particular ASR program and daily logs would help strengthen the claims.

References

Ahn, T.Y., and 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.

Chen, H. H.-J. (2011). Developing and evaluating an oral skills training website supported by automatic speech recognition technology. *ReCALL,* 23(1): 59–78.

Coniam, D. (1999). Voice recognition software accuracy with second language speakers of English. *SYSTEM,* 27(1): 49-64.

Cucchiarini, C., Neri, A., and Strik, H. (2009). Oral proficiency training in Dutch L2: The contribution of ASR-based corrective feedback. *Speech Communication,* 51(10), 853-863.

Cucchiarini, C., Strik, H. and Boves, L. (2000). Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithm. *Speech Communication,* 30(2–3): 109–119.

Dodd, S., and Mills, J. (1996). *Phonetics and phonology. Solving language problems: from general to applied linguistics*, 13–33. CORE: University of Exeter Press.

Derwing, T. M., Munro, M. J., and Carbonaro, M. (2000). Does popular speech recognition software work with ESL speech? *TESOL Quarterly,* (34): 592–603.

Ehsani, F., and Knodt, E. (1998). Speech technology in computer-aided language learning: Strengths and limitations of a new CALL paradigm. *Language Learning and Technology,* 2(1): 54–73.

Eskenazi, M. (1999). Using a computer in foreign language pronunciation training: What advantages*? CALICO Journal, (*16): 447–469.

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

Flege, J. E., (2007). Language contact in biliangualism: Phonetic system interactions. *Labarotory Phonology*, (9): 353-382.

Kirkova-Naskova, A. (2012). Interlanguage phonology: comparison between the English and the Macedonian vowel systems. In *Annual Symposium of the Faculty of Philology 'Blaze Koneski* (38): 141-152 [original] Киркова-Наскова, А. (2012). Меѓујазична фонологија: споредба на вокалните системи на англискиот и на македонскиот јазик. Во: *Годишен зборник на Филолошкиот факултет „Блаже Конески",* кн. 38, 141–152.

Kirkova-Naskova, A. (2010). Native speaker perceptions of accented speech: The English pronunciation of Macedonian EFL learners. *Research in Language,* (8): 1-21.

Levis, J., and Grant, L. (2003). Integrating pronunciation into ESL/EFL classrooms. *TESOL Journal,* 12(2): 13-19.

Levis, J., and Suvorov, R. (2013). Automatic speech recognition. In C. Chapelle (ed.). *The encyclopedia of applied linguistics*, 1–8. Hoboken, NJ: Blackwell Publishing.

Liakin, D., Cardoso, W., and Liakina, N. (2014). Learning L2 pronunciation with a mobile speech recognizer: French /y/. *CALICO Journal,* 32*(*1): 1-25.

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

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

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., Cucchiarini, C., and Strik, H. (2006). Selecting segmental errors in L2 Dutch for optimal pronunciation training. *International Review of Applied Linguistics, (*44): 357–404.

Neri, A., Cucchiarini, C., and Strik, H. (2008). The effectiveness of computer-based speech corrective feedback for improving segmental quality in L2 Dutch. *ReCALL: the Journal of EUROCALL*, *20*(2): 225.

Neri, A., Cucchiarini, C., Strik, H.Boves, L. (2002). The pedagogy-technology interface in Computer Assisted Pronunciation Training. *Computer Assisted Language Learning,* 15(5): 441–467.

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In *S.* Gass and C. Madden (eds.). *Input in second language acquisition*, 235–253. Rowley, MA: Newbury House.

Victori, M., and Lockhart, W. (1995). Enhancing metacognition in self-directed language learning. *System, 23*(2): 223-234.

McCrocklin, S. M. (2014). *The potential of Automatic Speech Recognition for fostering pronunciation learners' autonomy.* Graduate Theses and Dissertations. 13902. https://lib.dr.iastate.edu/etd/13902

Thomas, D.R. (2003). *A general inductive approach for qualitative data analysis.* [Online] Available from: www.health.auckland.ac.nz/hrmas/Inductive2003.pdf . [Accessed: May 5th, 2020]