Konin Language Studies
Faculty of Philology, State University of Applied Sciences in Konin, Poland
KSJ 7 (2). 2019. 137-157
http://ksj.pwsz.konin.edu.pl
doi: 10.30438/ksj.2019.7.2.2

Dictation programs for second language pronunciation learning: Perceptions of the transcript, strategy use and improvement

Shannon McCrocklin
Southern Illinois University, USA
https://orcid.org/0000-0002-5365-2067
shannon.mccrocklin@siu.edu

Abstract

Despite growing evidence that ASR-dictation practice provides benefits for L2 pronunciation learners (Liakin, Cardoso, & Liakina, 2014; McCrocklin, 2019; Mroz, 2018, Wallace, 2016), there is little research into the ways students engage in ASR-dictation practice. This study examines learners' perceptions of the ASR-generated transcript as feedback and strategy use during practice. Participants (*N* = 15) dictated 60 sentences to *Google Voice Typing* in *Drive* while being audio recorded. Following a mis-transcription, participants thought-aloud, discussing their interpretation of the transcript, utilized strategies and resources, and tried the sentence again with *Google.* Data analysis included qualitative analysis of think-aloud comments and quantitative analysis of both strategies used and improvement in dictation accuracy for subsequent attempts. Results showed that participants used the transcript to identify individual words with errors, but also hypothesized about segmentals and articulatory features causing errors. The most frequent strategy to improve production was covert rehearsal of target words, followed by listening to dictionary recordings of targets. Possibly novel pronunciation learning strategies were also documented, however. Participants were able to improve the accuracy of the transcript in subsequent attempts, earning a perfect transcription by the third attempt in the majority of cases (91%).

Keywords: pronunciation; learning strategies; Computer-Assisted Language Learning (CALL); Automatic Speech Recognition (ASR); English as a Second Language (ESL)

1. Introduction

In recent years, researchers have shown renewed interest in dictation programs, which use Automatic Speech Recognition (ASR) to provide a written transcript of speech, for second language (L2) pronunciation learning (Liakin, Cardoso, & Liakina, 2014, 2017; Mroz, 2018; Wallace, 2016). Despite growing evidence that ASR-dictation practice provides benefits for learners (Liakin et al. 2014, 2017; McCrocklin, 2016, 2019a), however, there is little research into how users engage with ASR-dictation programs as part of pronunciation practice. As interest in dictation programs for L2 pronunciation learning and teaching revives, research into learner perceptions of the dictation transcript as feedback, use of strategies and resources during practice, and ability to improve transcription accuracy through practice is needed.

2. Literature review

2.1. Dictation programs for pronunciation practice

Initial interest in dictation for L2 pronunciation followed important advancements in ASR technology in the 1980s and 1990s (Rabiner & Juang, 2008). Researchers grew interested in the potential of dictation programs for providing pronunciation feedback for English as a Second Language (ESL) learners. At the time, researchers found the programs to have inadequate recognition for L2 speech, preventing useful and reliable feedback (Coniam, 1999; Derwing, Munro, & Carbonaro, 2000). Coniam (1999) first raised concerns about the accuracy of dictation transcripts after finding substantial differences between the accuracy rates for native and non-native speakers in *Dragon Naturally Speaking,* an ASR-based dictation program commercially available through Nuance. Derwing, Munro and Carbonaro (2000) further established those concerns with an examination of 30 participants' (10 L1 English, 10 L1 Spanish, and 10 L1 Chinese) dictations of 60 sentences to *Dragon Naturally Speaking*, while also audio-recording each participant's speech. Then, they compared the program dictation accuracy of the 60 sentences (two selected from each speaker) to 41 native-speaker listeners who wrote what they heard (measure of intelligibility) and rated each speech sample on accentedness and comprehensibility. Finally, expert raters marked each sentence for phonemic errors. They found that *Dragon Naturally Speaking* did not perform as well as human listeners. While software recognition of native speech was 90.25% (versus 99.7% for human listeners), software recognition for non-native speech was 72.45% for Chinese first language (L1) speakers and 70.75% for Spanish L1 speakers (compared to 94.99%

and 95.71% for human listeners, respectively). The program's transcription rates also did not have any statistically significant correlations with the speakers' intelligibility, comprehensibility, or accentedness as decided by the human listeners or the phonemic accuracy as marked by expert raters. Strik, Neri and Cucchiarini (2008) raised additional questions about the usability of feedback provided by dictation output, arguing that "dictation programs are not suitable for L2 training, CALL requires dedicated speech technology" (p. 74). The field moved on to focus on CAPT (computer-assisted pronunciation training) programs that could embed ASR in order to provide more explicit feedback for learners on preprogrammed words and phrases (Cucchiarini & Strik, 2018). Research into dictation programs paused until recently.

Although more research is needed to examine current dictation accuracy, recent research has shown that dictation programs are beginning to show improvements for non-native speech. McCrocklin, Humaidan and Edalatishams (2019) compared the accuracy rates of *Google* and *Windows Speech Recognition*, finding that while *Windows Speech Recognition* showed no improvements from the reported accuracy rates in Derwing et al. (2000), *Google* had noticeably improved its transcription for non-native speech to 91.04% (average across two different tasks) compared to 94.98% for native speakers. More research comparing the transcript accuracy to human listeners is needed, but as ASR programs have improved, researchers have taken notice and recent research shows benefits of ASR-dictation practice.

2.2. Benefits of dictation practice

When compared to CAPT, dictation programs have two main advantages: greater accessibility and flexibility. Unlike many of the CAPT programs that use ASR, many ASR-dictation programs, such as *Windows Speech Recognition* and *Mac Dictation*, are available as accessibility services for no additional charge as part of computer operating systems. Further, ASR-dictation through *Google*, such as *Voice Typing* in *Drive,* or voice searching are available for free when browsing in *Chrome*. Many students and teachers already own a device that has, or can easily access, a freely-available ASR dictation program. Once students know how to access dictation services, many find the programs easy to use (McCrocklin, 2019b; Mroz, 2018). Finally, dictation programs are flexible; they can work to dictate any content into a text form, allowing students to choose content and direct their work. Teachers can use this flexibility to integrate ASR-dictation practice more effectively into their courses (McCrocklin, 2015).

Further, dictation practice can provide a range of benefits including noticing of pronunciation issues (McCrocklin, 2019b; Wallace, 2016), increased motivation

and autonomy (McCrocklin, 2016, Mroz, 2018), and improvement in segmental accuracy (Liakin et al., 2014; McCrocklin, 2019a). Wallace (2016) first argued that dictation programs may be useful for raising awareness of pronunciation issues. McCrocklin (2019b) supported this, finding that participants reported that one of the main advantages of using ASR-dictation practice was noticing errors and gaining a heightened awareness of their personal pronunciation weaknesses. Further, exposure to and practice with ASR-dictation programs led participants to report greater learner autonomy in regards to pronunciation practice in McCrocklin (2016), and participants reported greater motivation for learning as they gained insight into the ways they may be understood by human listeners in Mroz (2018). Perhaps most importantly, practice with ASR dictation can help learners improve their segmental production. Liakin et al. (2014) examined participant productions of the French vowel /y/ using a pre-/post-test design. When comparing three groups (an ASR-dictation practice group, a non-ASR pronunciation training group, and a control group with no training), only the ASR-dictation group made statistically significant improvements in the French vowel /y/. McCrocklin (2019a) also examined student improvement using listener ratings of accuracy for several targeted sounds (English consonants and vowels) in a pre-/post-test design. Participants in a workshop using ASR-dictation for half of their production practice improved as well as the entirely face-to-face instruction group, slightly outperforming the face-to-face group on most segmentals.

2.3. Student experience and perceptions of ASR-dictation practice

Though dictation practice has received somewhat mixed reviews from participants, researchers tend to conclude that advantages outweigh drawbacks (Liakin et al., 2017; Mroz, 2018). The majority of studies on dictation programs, however, investigate the impact of ASR-dictation after a particular type of training or practice has concluded. One of the findings of these studies is that participants sometimes report frustration as they work with ASR-dictation programs (Liakin et al., 2017; McCrocklin, 2019b). Although both studies suggested that a lack of accuracy in the dictation transcript may have prompted feelings of frustration, it was unclear to what degree the transcripts may have been helpfully indicating pronunciation errors. Further, as students receive no explicit feedback from the program, frustration may have emerged due to participants' uncertainty about how to improve their pronunciation in subsequent attempts. While Strik et al. (2008) questioned whether the transcript could be considered usable feedback, McCrocklin (2019b) reported that many participants believed that they did receive usable feedback from the dictation transcript. It is unclear, however, to what degree students can glean specific feedback from a transcript. More information about sources of possible frustration,

including possible lack of improvement in subsequent attempts or lack of clear feedback from the transcript, need further exploration.

Previous studies also raised questions about how strategies and resources may be used in practice with ASR-dictation. Use of strategies can positively impact pronunciation improvement and ultimate attainment (Moyer, 2014). Earlier research by Osburne (2003) showed that when prompted to improve pronunciation in a subsequent attempt with an interlocutor without any feedback about where a miscommunication occurred, participants reported strategies such as imitation of an interlocutor, attention to individual words and attention to paralanguage features. One study examining strategy use with CAPT programs, Fang and Lin (2012), found that similar strategies were used in CAPT and face-to-face instruction, including frequent use of mimicry of a provided model and focus on paralanguage while rarely attending to segmental issues. ASR-dictation may also lead to high rates of mimicry, as McCrocklin (2019b) found that participants reported using primarily e-dictionaries to listen to target words, although covert rehearsal (private practice of the sound, word, or sentence for self-monitoring) was also mentioned. However, there are no studies examining student perceptions and strategy use during ASR-dictation practice. Pawlak and Szyszka (2018) call for more research into pronunciation learning strategies employed during different tasks, as they likely differ.

The current study explores participants' mental processes and practice patterns while using *Google Voice Typing* for pronunciation practice. In particular, it addresses three research questions:

1. How do participants make sense of the provided transcript as feedback on their pronunciation?
2. What resources and strategies do participants make use of in their practice with dictation programs and what is the relative frequency?
3. To what degree can participants improve the accuracy of transcription provided by *Google's* ASR in subsequent attempts?

3. Methodology

The current research study examined student interpretations of the transcript, strategy use, and transcription accuracy changes through a mixed-methods design, which included: qualitative analysis of participants' think-aloud comments regarding their perceptions of the transcript as feedback, quantitative analysis including counts of resources and strategies employed before a subsequent attempt, and quantitative analysis of dictation transcript accuracy across multiple attempts. More information about the participants, procedure, and analysis are provided in the following subsections.

3.1. Participants

Participants (*N* = 15) were undergraduate and graduate students at a mid-sized university in the United States. Participants spoke a variety of native languages: Chinese (*N* = 7), Spanish (*N* = 5), Arabic (*N* = 1), Japanese (*N* = 1), and Ambonese/Malay-Indonesian (*N* = 1). Participants reported an average age of 25.8 (*SD* = 7.20), had spent an average of 17.7 years learning English (*SD* = 6.05), and had lived in the U.S. an average of 1.9 years (*SD* = 2.93). Participants self-reported their TOEFL score, which averaged 87.4 on the IBT (*SD* = 11.91). The participants were split by gender; 53.33% were male, while 46.47% were female. In addition to learning English, the majority of participants (*N* = 11) reported having studied or learned a third language.

3.2. Procedure

Participants scheduled a one-hour time slot in which to participate in the study in a lab on campus. Upon arriving, participants were introduced to the study and then provided informed consent. Participants also answered demographic questions in a short questionnaire. They were then introduced to the task, including directions for the think-aloud protocol with example questions and issues to consider, and resources available to help them with their pronunciation if desired. These resources included *Dictionary.com* (which includes audio recordings for each word), *Soundsofspeech.uiowa.edu* (which includes animations and audio samples of English segmentals), and *Youglish.com*. Each of the web-based resources was left open in a tab on the browser. A final resource introduced was mini-lessons from the researcher. Piloting of the study showed that participants occasionally wanted to be able to access lessons about the articulation of sounds or words, but videos accessed on *YouTube.com* were often too lengthy to be integrated efficiently into the ASR practice. Thus, the researcher, when prompted, provided 1-2 sentences with articulation information on how sounds/words differed or articulation tips for creating the desired sound/word. Participants were invited to access additional resources on the internet or use additional strategies as desired.

Then, participants dictated 60 sentences (controlled, read speech) to a document in *Google Drive* using *Google's Voice Typing*. The sentences were a mixture of true/false sentences that had an average sentence length of 6.1 words and featured a variety of non-technical vocabulary. The sentences were similar in format and content to those used in Derwing et al. (2000). In the case that the dictation program mis-transcribed the sentence,[1] participants were

[1] In the case of a homophone, utterances were counted as correct and the participant was encouraged to move to the next sentence. Only one homophone emerged in the data, the possessive "grandmother's" which was consistently transcribed in place of the plural "grandmothers".

prompted to discuss what feedback, if any, they perceived in the transcript. The piloting of the study showed that participants tended to skip the think-aloud protocol if the researcher did not prompt. Thus, prompts were provided frequently to ensure sufficient data was collected about participants' perceptions of the transcript. After thinking aloud, participants chose and utilized resources or strategies as desired before re-recording any mis-transcribed sentences. This process repeated until *Google* produced an accurate transcript of the target sentence or until the participant had tried the sentence four times. At four attempts, the participant was encouraged to move on to the next sentence. On average, the task of reading the 60 sentences, including making think-aloud comments, utilizing resources and strategies, and repetitions in subsequent attempts, took 35 minutes (range: 15-49.5 minutes). The dictation practice and think-alouds were audio-recorded using a Logitech USB microphone and *Audacity*. The dictation output was saved in the *Google Drive* document. Participants' information was saved using a random numerical identifier (P1, P2, etc.) and all data was saved under this identifier.

3.3. Analysis

The analysis included several different steps. To address the first research question regarding participants' understanding of the transcript as feedback, themes were identified from comments elicited through the think-aloud protocols during practice which were transcribed verbatim and coded based on areas of participant focus. Osburne (2003) was used to create a starting list of possible themes to examine (foci on articulatory gesture, single sound, individual syllable, and prosodic structure), but the transcript was also explored for additional possible themes or strategies using a general inductive approach. To address the second research question, the analysis also included a quantitative analysis using descriptive statistics to examine participants' use of resources and strategies before a subsequent attempt. The researcher noted use of strategies and resources while the participant worked on the task. All strategies and resources were accompanied with clear auditory cues (speaking a word outloud, listening to a word in the dictionary, etc). Using the audio recordings, each use of a resource or strategy was later counted, including how many times each strategy was utilized (with the exception of the articulatory mini-lessons from the researcher which were often hard to quantify and therefore counted as a single use once activated). To address the final research question, the transcript provided by *Google* was examined for accuracy. For each sentence, the percent of accurate transcribed words from the original sentence was calculated. For example, for the sentence "Some people think knitting is relaxing", one participant

received the transcript: "Some people think anything is relaxing", which was given a score of 5 out of 6, for 83.33% correct. Each word was worth equivalent value in the analysis. In the case that *Google* recognized the stem of a word but included a morphological error, the transcribed word was assigned half credit. For example, for the sentence "The postal worker makes deliveries to your home", another participant received the transcript: "The postal worker makes delivers to your home", which received a score of 7.5 out of 8, for 93.75% correct. For mis-transcribed sentences, accuracy was also calculated for each subsequent attempt.

4. Results

Of the initial 900 sentences (15 participants reading 60 sentences), 383 featured errors in the first attempt, thus prompting a second attempt. A little under half of the 383 sentences (N = 179) required a third attempt and only about half of those (N = 82) were taken to a fourth attempt. It is important to note that participants may not have acquired a perfect transcript in the fourth attempt; in fact, only a quarter (24.39%) of the 82 sentences were transcribed correctly in the fourth round. Participants were encouraged to move on to the next sentence at that point, however. When totaled, the 383 sentences, tracked through multiple attempts, led to 644 instances in which a sentence needed to be repeated in a subsequent attempt.

4.1. Participants' perceptions of provided transcript as feedback on pronunciation

The analysis of participants' comments during think-aloud protocols (made while working to make sense of the dictation transcript as potential feedback on their pronunciation) identified six major themes. Five of the themes represent hypotheses that participants created based on the transcript and generally align with areas of focus identified in Osburne (2003), while the final theme, "Uncertainty/Questions" indicate instances that a participant asked a question about their pronunciation, doubted the transcript, or was uncertain of the feedback that could be gleaned from the transcript. Table 1 provides the number counts and percentages of a particular focus across each of the attempts along with the total number of participants that reported that focus across all attempts.

The analysis showed that out of the 644 instances in which participants did not receive a perfect transcription on an attempt, participants most frequently commented on at least one word that they noticed or believed may have contained pronunciation issues based on their perception of the transcript (30.90% of cases). All participants commented about an individual word that caught their attention at least once during their practice. For example, when P1 tried the sentence "All pens have purple ink", it was transcribed as, "Opens have

purple ink". In thinking aloud, he said "I have an issue with *all* I think". The second most common area of focus was a consonant or vowel (segmental) that they hypothesized was pronounced incorrectly (12.73%). Almost all participants (93.33%) hypothesized about a potential segmental issue at some point in their practice. These guesses were often reasonable given the transcript. For example, when P5 tried the sentence "Some children ride their dogs to school", which was transcribed as "Some children write their jaws to school", she said, "I think it's a /d/ sound probably". In only a small amount of cases, did the participants focus on an articulatory feature/gesture or a suprasegmental feature (4.04% and 1.71% respectively). However, it was often unclear in these cases if participants used the sometimes technical vocabulary appropriately in their explanations. For example, P10 after dictating the sentence "Cheetahs run very slow", and receiving the transcript, "Cheetahs run ferry slow", he said "Oh again. What is happening with my /f/? Um… I say this is the most difficult part for me to distinguish /f/ and /v/. They are very similar. Interdental. The stress to /v/ which one is the sharpest. They are very similar". While the term *interdental*, was very close to an appropriate term, such as *labio-dental* to describe the similarities between /f/ and /v/, it is unclear what the participant might have meant by the term, *stress*. Finally, in a small number of cases, participants noted speed as a potential issue preventing correct transcription (2.02% of cases). For example, when P15 tried the sentence "People can look up the date in a calendar", but *Google* transcribed "People can look up the dating a calendar", she remarked, "maybe too fast". In most of those comments, participants hypothesized that they had spoken too quickly, although one participant hypothesized that they needed to increase their speed in the subsequent attempt.

Table 1 Themes identifying participant focus in think-aloud comments by participant and attempt (with counts of themes and percentages out of total number possible)

	Post-attempt 1		Post-attempt 2		Post-attempt 3		Total across all attempts		Participants across all attempts	
Total possible		383		179		82		644		15
Focus	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Segmental	49	12.79	24	13.41	9	10.98	82	12.73	14	93.33
Suprasegmental	6	1.57	3	1.68	2	2.44	11	1.71	7	46.67
Word as a whole	136	35.51	51	28.49	12	14.63	199	30.90	15	100.00
Articulatory feature	8	2.09	15	8.38	3	3.66	26	4.04	10	66.67
Speed	10	2.62	2	1.12	1	1.22	13	2.02	8	53.33
Uncertainty/Questions	62	16.19	22	12.29	12	14.63	96	14.91	15	100.00

Note. For an explanation of the number possible for each post-attempt, please see the first paragraph of the results section.

Participants voiced uncertainty about the cause only 96 times (14.91%). Further, only 14 of those (2.17% of the total 644 cases) were instances in which the participant voiced uncertainty without indicating a specific word that caught their attention or without proposing some form of hypothesis about what might be happening in the transcript, suggesting that participants are rarely at a complete loss about the source of their pronunciation error. Even in cases that participants did not have a specific hypothesis, they often turned to the dictionary for the word that they noticed as wrong in the transcript. For example, when P14 read the sentence “All cats have black paws”, and received the transcript “All cats have black balls”, she said “What? I'm making a /p/ sound. I don't know what I'm doing wrong! I know how to pronounce it!” Although at first she was frustrated and did not know what may have caused the transcription error, she chose to listen to the dictionary entry for “paw” twice. At this point, she asked with some surprise, “Is that how you pronounce it?” She practiced the word aloud and self-monitored (covert rehearsal) three times before trying the sentence again and received a perfect transcription on the subsequent attempt.

Examining trends across attempts, participants grew less likely to comment on the word as a whole in subsequent attempts. It is important to note, however, that participants often received transcripts with errors on the same words identified in previous attempts and likely continued to be aware of the word in subsequent attempts. From the second to third attempt, participants became more focused on articulatory features. For the sentences that were only taken to a third attempt, participants also grew more focused on segmentals moving into the third attempt, which brought up the average for post-attempt 2 slightly. This may suggest that with additional feedback from the transcript, participants were able to focus more on the specific parts of words that may have caused a transcription error. Additionally, an interesting trend was that for those who needed a third attempt, they mentioned less uncertainty moving into the third attempt, but for the sentences that needed a fourth attempt, uncertainty remained stable across all attempts. This may suggest that sentences that still feature errors following a third attempt may need intervention in order to help participants recognize issues present in their production.

As a follow-up analysis, the transcripts were checked to see if participants recognized a word that was repeatedly mis-transcribed or noticed a pattern of sounds that were frequently mis-transcribed. Every participant remarked upon some form of pattern of errors; 12 of the 15 participants noted a word that showed repeated transcription errors, while 6 of the 15 noticed a pattern of segmentals that they hypothesized were creating errors in transcription across multiple words.

4.2. Resources and strategies participants make use of in their practice

After thinking aloud, participants were encouraged to make use of online resources or additional strategies to improve their pronunciation before trying the sentence again. The analysis of strategies and resources utilized by participants identified nine different strategies (see Table 2). The strategies were primarily cognitive strategies, but some also included aspects of metacognitive and social strategies. Four of the resources and strategies were those introduced to participants, but an additional five strategies were added through the analysis. Most of the strategies aligned with previously identified pronunciation strategies in Peterson (2000), which included talking aloud to oneself (cognitive), listening to tapes (cognitive), deciding to focus one's listening on particular sounds (metacognitive), and asking someone else to correct one's pronunciation (social).

Table 2 Strategies and resources documented during ASR-dictation practice

Resource/Strategy	Explanation	Strategy group (Peterson, 2000)
Covert Rehearsal (CR) Target	covert rehearsal, practice individual words aloud for self-monitoring	Cognitive
CR Phrase	covert rehearsal, practice two or more words of the sentence aloud for self-monitoring	Cognitive
CR Transcribed	covert rehearsal, practice transcribed word aloud for self-monitoring	Cognitive
Dict. Listen Target	using the dictionary to look up and listen to the target word	Cognitive
Dict. Listen Transcribed	using the dictionary to listen to the transcribed word	Cognitive
Youglish	using Youglish.com to listen to the target word	Cognitive
Articulation Mini	requesting a mini-lesson of articulatory information, including potentially tips to make an intended sound or explaining pronunciation differences between two words	Metacognitive and social
Iowa SS	using Iowa Speech Sounds to listen to a challenging phoneme in a target word	Cognitive and metacognitive
ASR Target	practice individual word with ASR-dictation before trying it back in the phrase	Cognitive

Table 3 Use of strategies/resources following a transcription error

	Ave use post-attempt 1		Ave use post-attempt 2		Ave use post-attempt 3		Ave total use	Used by # of participants
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*n*
CR Target	1.73	2.50	1.70	2.46	1.17	1.78	1.65	15.00
CR Transcribed	0.40	0.81	0.36	0.75	0.17	0.47	0.36	15.00
CR Phrase	0.24	0.61	0.28	0.96	0.20	0.73	0.24	14.00
Dict. Listen Target	0.59	1.07	0.69	1.13	0.49	0.79	0.61	15.00
Dict. Listen Transcribed	0.04	0.22	0.08	0.35	0.12	0.53	0.06	12.00
Youglish	0.02	0.21	0.07	0.50	0.18	0.67	0.05	5.00
Articulation Mini	0.14	0.35	0.18	0.40	0.17	0.38	0.16	12.00
Iowa SS	0.00	0.11	0.11	0.73	0.11	0.80	0.05	6.00
ASR Target	0.00	0.00	0.03	0.38	0.00	0.00	0.01	2.00
Sum Total	3.16		3.50		2.61		3.19	

Each use of a resource or strategy was counted. Because participants could engage multiple strategies or could re-use a strategy multiple times, it was important to track how many times each strategy was used on average between attempts. Table 3 provides the average number of uses of each strategy and resource following each attempt along with counts of the number of participants that used each of the strategies or resources across the entire practice session.

On average, a participant made use of 3.19 strategies or resources before a subsequent attempt. The analysis shows that the most frequently utilized strategy was covert rehearsal of a single target word. On average, participants rehearsed a selected target word 1.65 times before trying a subsequent attempt. The second most commonly used strategy was listening to the target word in the dictionary. Participants listened to the word in the dictionary, on average, 0.61 times before a subsequent attempt. Covert rehearsal also made up the third and fourth most common strategies employed; CR Transcribed was employed 0.36 times before a subsequent attempt while CR Phrase was employed 0.24 times. The fifth most common strategy was to request an articulation mini-lesson (employed 0.16 times before a subsequent attempt). Notably, other strategies could be employed and counted multiple times (for example, listening to the dictionary recording five times before trying again would count as five uses), but asking for articulation advice, even if it included multiple tips or there were follow-up questions only counted for a single activation because of the challenges of counting the number of tips. This brought the average use down despite it being frequently employed overall. The least common strategies were Youglish (0.05 times), Dict. Listen Transcribed (0.06 times), Iowa Speech Sounds (0.05 times), and ASR Target (0.01 times).

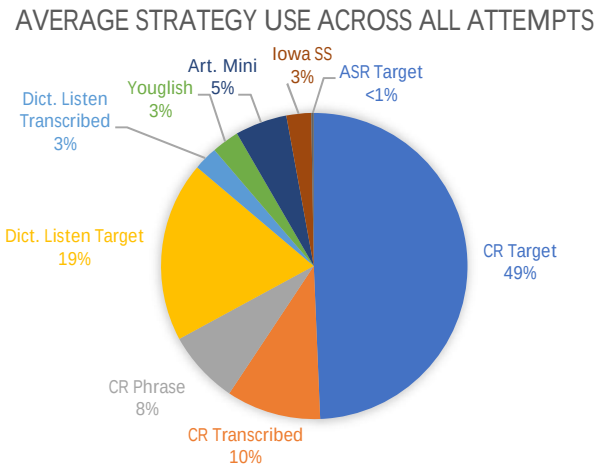

Figure 1 Frequency of use of strategies/resources in ASR-dictation practice

Figure 1 shows a visual display of the relative frequency of use of each of the strategies as a percentage of all strategies and resources used. Figure 1 highlights that the majority (68%) of the strategy use focused on the target word.

The analysis provided in Table 3 also shows that participants used the most strategies between the second and the third attempt (average of 3.50 strategies) while they used the least strategies between the third and fourth attempt (average of 2.61 strategies). The lower strategy usage moving into the fourth attempt may indicate that participants were growing frustrated or fatigued or that they were not sure how to continue to use resources to improve. Strategy use following the first attempt averaged 3.16 strategies before the subsequent attempt.

As a follow-up, a brief analysis of participant comments during the think aloud which followed the first attempt was conducted, which suggests that sometimes participants doubted there was a pronunciation issue following the first attempt, but realized there likely was a pronunciation problem upon confirmation of the error in the second attempt. The confirmation then prompted additional resource usage. For example, when P11 tried the sentence "Cats often bark at strangers", it was transcribed as "Hats off and bark at strangers". P11 decided to simply try it again with no think aloud and no use of strategies or resources. After getting the same transcription again, he said "I might have an issue with the word *often*. Yeah, it's probably the way that I'm pronouncing the initial vowel so I want to try to repeat this one". He then tried listening to the word in the dictionary twice and tried covert rehearsal of the word once and achieved a perfect transcription on the third attempt. Occasionally, participants also indicated that they believed they could self-correct without further intervention following a first attempt and moved straight to a second attempt. However, without more questioning directed at participants on this specific issue it was difficult to draw firm conclusions.

4.3. Participants' ability to make changes and improve accuracy of transcription

The accuracy of the transcript provided by *Google* for each sentence was analyzed by counting the number of correct words from the original sentence included in the transcript. This process was repeated for all subsequent attempts (see Table 4).

The results show that for sentences that did not result in perfect transcriptions, participants were able to improve the transcript accuracy on the second and third attempts. Participants improved accuracy by around twelve percent for each repetition: 12.65% increase in accuracy in second attempt for repeated sentences and 11.59% increase in accuracy in third attempt for repeated sentences. Notably, while participants were able to continue making improvements in the second and third attempts, the fourth attempt did not show similar improvements. For

sentences that did not achieve a perfect score in the third attempt, the average accuracy was 78.41% while those sentences in the fourth attempt dropped to 76.84% accurate. Figure 2 shows a graphical representation of this relationship. For each attempt, the figure shows the accuracy for all sentences in the attempt (the higher score) and the score for the sentences that needed to be repeated, which connects in a line to the accuracy for those same sentences in the next attempt.

Table 4 Accuracy of transcripts for each attempt, number of sentences that required a subsequent attempt, and accuracy of sentences in each attempt that moved on to subsequent attempt

Participant	1st Attempt accuracy	# Sentences to be repeated in 2nd attempt	1st Attempt accuracy (for only repeated sentences)	2nd Attempt accuracy	# Sentences to be repeated in 3rd attempt	2nd Attempt accuracy (for only repeated sentences)	3rd Attempt accuracy	# Sentences to be repeated in 4th attempt	3rd Attempt accuracy (for only repeated sentences)	4th Attempt accuracy
	M	*n*	*M*	*M*	*n*	*M*	*M*	*n*	*M*	*M*
P1	87.94	27.00	69.05	87.00	5.00	73.24	90.00	2.00	81.67	70.00
P2	83.33	31.00	67.96	75.96	18.00	67.41	82.34	3.00	85.69	84.58
P3	91.04	22.00	75.57	86.64	12.00	76.63	85.87	4.00	74.76	77.26
P4	85.30	34.00	74.06	88.31	18.00	77.91	93.96	7.00	84.48	84.15
P5	87.57	32.00	76.69	88.24	17.00	78.61	91.90	8.00	82.78	77.74
P6	88.38	32.00	78.22	90.58	13.00	77.63	79.66	10.00	74.53	75.19
P7	88.10	27.00	73.55	87.77	15.00	79.32	88.45	8.00	78.35	79.42
P8	83.58	37.00	74.05	88.04	16.00	74.71	85.07	9.00	70.93	81.24
P9	86.43	30.00	72.85	85.46	15.00	70.92	83.93	10.00	75.89	77.33
P10	94.63	17.00	81.06	92.30	6.00	78.17	91.58	3.00	83.15	91.07
P11	91.06	21.00	74.47	89.27	9.00	74.97	88.41	4.00	73.93	78.93
P12	85.94	29.00	70.92	86.06	17.00	76.22	85.37	8.00	68.91	76.01
P13	93.62	19.00	79.86	88.62	8.00	74.40	95.71	2.00	82.86	90.00
P14	96.44	11.00	80.60	89.77	6.00	81.25	96.67	1.00	80.00	20.00
P15	95.31	14.00	79.91	94.57	4.00	87.26	83.69	3.00	78.25	89.68
M	89.25	25.53	75.26	87.91	11.93	76.58	88.17	5.47	78.41	76.84
SD	4.24	7.77	4.13	4.07	5.12	4.56	5.06	3.18	5.11	16.83

Note. Columns 4, 7, and 10 show the accuracy of for the attempt after removing perfect scores to allow for comparison with the scores achieved in the subsequent attempt among the smaller subsection of sentences.

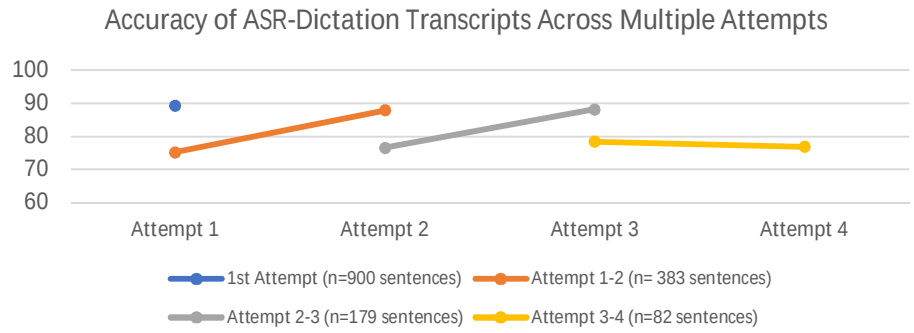

Figure 2 Accuracy of dictation transcript across multiple attempts organized by sets of repeated sentences

To check whether a small set of sentences provided an outsized number of the attempts, a follow-up analysis was conducted to explore if there were patterns in the sentences that were most challenging, those that needed a fourth attempt. The analysis showed that 33 different sentences needed a fourth attempt by at least one participant. The two most challenging sentences each needed a fourth attempt by six different participants. Those sentences were "All pens have purple ink (with most transcription errors on the words "all" and "pens") and "At the theater you can see many plays (with the most transcription errors on the word "plays"). Upon analysis of these sentences, participants tended to feature pronunciation errors that could account for the transcription errors in these sentences, including vowel quality issues in "all" and "pens" and the devoicing of the /z/ in the plural of "pens" and "plays". The fact that there are some sentences that never made it into a fourth attempt and some sentences that numerous participants took to a fourth attempt, however, suggests that the sentences themselves also played a role in the ability of participants to improve transcript accuracy in subsequent attempts.

5. Discussion

While recent research has shown several benefits of ASR-dictation practice, including noticing of pronunciation issues (McCrocklin, 2019b; Wallace, 2016), increased motivation and autonomy (McCrocklin, 2016; Mroz, 2018), and improvement in segmental accuracy (Liakin et al., 2014; McCrocklin, 2019a), no studies had previously examined participant perceptions, strategy use, and improvement during ASR-dictation practice. Using think aloud protocols, tracking of strategy use, and analysis of transcript accuracy, several important findings emerged.

Similar to previous work by Osburne (2003), participants most commonly used the transcript to focus on individual words that had been mis-transcribed. However, the results differed from previous research in that participants were sometimes able to further hypothesize about specific segmental or articulatory features of their speech that may have contributed to the transcription error, suggesting ASR-dictation practice may be particularly useful in prompting participants to think about sounds and articulatory features. While participants did occasionally voice confusion or uncertainty about the feedback, the number of cases in which participants had no hypothesis about the source of an error was very small suggesting ASR-dictation transcripts may be useable as feedback on pronunciation if the transcript accurately reflects pronunciation errors, which is an area of work that is needed in future studies.

While some of the findings regarding participants' focus in interpreting the feedback align with previous studies (Osburne, 2003; Pawlak & Szyszka,

2018), such as frequent focus on the word level, implementation of speed as a possible strategy, and use of mimicking (in this study using a dictionary recording) as common strategies, there were some differences. Speed is brought up by Osburne (2003) under the category of paralanguage, which was found to be a frequently reported strategy, including additional aspects such as voice dynamics and fluency. While participants in this study did mention speed, they did not mention any of the other possible paralanguage features in their comments. It may be that participants did not believe that the program was affected by elements of voice dynamics or that the read speech task prevented some dis-fluencies, such as fillers. Further, the ASR-dictation task elicited a high degree of comments about segmentals. This may indicate that the ASR-dictation practice is useful for prompting students to think about their segmental production. Given the relative lack of emphasis on segmentals found in Osburne (2003), the results may offer additional evidence that the transcript can be used as feedback to help students begin to identify segmental issues in their production. Further, it suggests that ASR-dictation practice may differ from CAPT practice (as documented in Fang & Lin, 2012) in encouraging participants to think about segmentals and articulatory differences.

Although Strik et al. (2008) raised concerns about the usability of a dictation transcript as feedback on pronunciation, McCrocklin (2019b) found that participants reported feeling not only that they were able to understand feedback from the transcript, but that the feedback helped them identify specific weaknesses in their pronunciation. Although it is outside of the scope of this paper to provide phonetic analyses to ensure accuracy of participant hypotheses, the results showed that the ASR transcript was particularly useful in helping participants identify words that may have had pronunciation issues. Participants were also able, however, to hypothesize in about 18.48% of cases a segmental, suprasegmental, or articulatory feature that may have led to the error in transcription. Further, every participant noticed a pattern of errors during practice and 40% of participants were able to link errors to a particular segmental occurring across multiple words. This noticing occurred from a single practice with ASR-dictation that lasted on average only 35 minutes. This supports McCrocklin's (2019b) findings that dictation practice can be useful for student identification of not only errors within a particular word or sentence, but also areas of weakness based on patterns of errors.

As participants prepared to try a sentence again in a subsequent attempt, they were most likely to say the words aloud as part of self-monitoring (covert rehearsal) or to use a dictionary to look up the target word. Although McCrocklin (2019b) noted several participants reported using e-dictionaries frequently in their L2 pronunciation practice with dictation, while only occasionally reporting

covert rehearsal, in this research study, covert rehearsal was the most frequently used strategy, making up almost half of all strategy and resource use. This difference demonstrates the importance of investigation of participants' perceptions and strategy use during practice with think-aloud protocols and careful tracking of strategy and resource use. Further, the analysis noted strategy use not previously reported as part of dictation practice, including attention to the transcribed word, both through covert rehearsal and listening to the dictionary, as participants worked to discover the differences between the words. The analysis also identified requests for articulation mini-lessons, which suggest that ASR-dictation practice may be useful for starting a conversation about pronunciation with an instructor or speakers of the language.

While the frequent use of covert rehearsal of a single target word is in line with several previous studies (Calka, 2011; Osburne, 2003), three of the resources and strategies used, may not fully align with tactics documented in previous research: listening to a dictionary entry of the transcribed word, covert rehearsal of the transcribed word, and ASR-dictation practice for the target word. Listening to the dictionary entry and using covert rehearsal of the transcribed word may partially fit under certain previously identified tactics, such as talking aloud to oneself, but also brings in an element of contrast that does not seem to be fully captured by the tactic of noticing contrasts between native and target language pronunciation. It also seems somewhat distinct from the strategy of minimal pair drills reported in Fang and Lin (2012) because the participant is not practicing with sets of minimal pairs to practice a contrast; instead, they are only focusing on the differences between the transcribed and target words, which notably may or may not be minimal pairs. Additionally, the use of the ASR-dictation program to check on an individual word and receive feedback is distinct from simply talking to oneself aloud and/or asking a person to correct one's pronunciation. No other strategies listed in Peterson (2000) fully capture the implementation of technology to get feedback.

Finally, the results show that participants were able to change their production in order to improve the accuracy of transcription in subsequent attempts. For the majority (78%) of sentences that did not receive a perfect transcription on the first attempt, participants were able to achieve a perfect transcription in the subsequent second or third attempt. Previous research has indicated that students can grow frustrated from a lack of recognition, particularly if students had to try numerous times (Liakin et al. 2017; McCrocklin, 2019b). One approach to limiting frustration could be providing guidance on how many times to try an utterance before moving on. Wallace (2016) asked students to try a passage only once while McCrocklin (2016, 2019a, 2019b) provided guidance to try a practice item up to three times before moving on. However, it was

unclear if students could make immediate modifications to improve their accuracy or how many attempts would be beneficial. Based on the results of this study, three attempts seemed to secure the maximum opportunities to improve the transcription accuracy. Participants had achieved a perfect transcription by the third attempt in 91% of cases. Further, participants failed to make substantial further progress in the fourth attempt, suggesting that the problems in those sentences may have needed more intensive intervention or that certain sentences may have been particularly challenging.

6. Limitations

Although this study has yielded interesting insight into the ways that participants interact with dictation programs as part of pronunciation practice, the findings are limited without a phonetic analysis of the pronunciation errors. Future analyses comparing mis-transcriptions to pronunciation errors are necessary. While participants could easily identify mis-transcribed words from the sentence transcripts, students could be led astray if they work from transcripts that do not reasonably represent the students' pronunciation. Although McCrocklin et al. (2019) showed notable improvements in the accuracy of transcription for *Google Voice Typing*, more research is needed to compare how transcripts provided to students resemble the intelligibility of human listeners. Further, future studies should include a wider variety of learners, including more language backgrounds and skill levels, as well as a variety of dictation programs to enable broader recommendations. It also might be helpful to track more than four attempts in a future study to see what participants are able to accomplish with additional successive attempts. Finally, future research should explore additional methods to investigate thoughts or behaviors during student practice with dictation programs, preferably including less intrusive methods than think-alouds. Because the researcher was constantly present, and frequently prompted the think-aloud behaviors, the practice with ASR-dictation was likely altered by participants' awareness of the researcher's presence as well as the constant pressure to think carefully about what the transcript may mean as feedback. Researchers could consider using audio recording and screen capture to remove some of the influence of the researcher.

7. Implications for teachers

Pronunciation instruction is valuable, even for beginning and intermediate learners (Zielinski & Yates, 2014). The findings suggest that participants are generally able to perceive feedback from the transcript, particularly focusing on mis-

transcribed individual words. ASR-dictation could be a valuable way to incorporate pronunciation instruction as part of early vocabulary learning. Students may notice incorrect pronunciations of new words early in learning and avoid issues in fossilization if ASR-dictation practice is incorporated. One possible way this could be implemented is an assignment in which students create sentences with new vocabulary they are learning and test out the new sentences with dictation. Further, teachers should encourage students to reach out to and use a variety of resources. McCrocklin (2019b) found that students valued e-dictionaries because of the recordings of words, which was supported in this study as well. Participants also valued practicing the target words, and even the transcribed words, while self-monitoring. However, additional resources may be useful. A few participants grew to enjoy *Youglish*, but almost none wanted to use the *Iowa Speech Sounds* website. They found it difficult to know how to find resources on the site, as sounds are categorized by their articulatory features and often asked for an articulatory mini-lesson when they wanted to know more about a specific sound. A teacher wishing to introduce *Iowa Speech Sounds* or articulatory mini-lessons would likely need to have guides for student practice with targeted sounds; these guides could then direct students to appropriate pages with links so that students are not required to navigate the site's confusing organization. A teacher could also find and link suitable videos online to replace some of the in-person articulatory lessons or could record their own. Additional resources should be explored, however, and may prove to be useful in the students' and teachers' repertoires to facilitate students' autonomous practice with dictation programs.

8. Conclusion

This study offers insight into the potential ways that students may interact with and perceive feedback from ASR-based dictation practice in a second language, as well as the ability to improve transcription accuracy on subsequent attempts. As the field of second language teaching rekindles interest in ASR-dictation practice and as ASR technologies evolve and advance, there is need for additional research into L2 pronunciation practice with dictation. In particular, future research should examine accuracy rates of ASR-dictation programs for non-native speech as well as student pronunciation improvements obtainable through ASR-dictation practice.

References

Całka, A. (2011). Pronunciation learning strategies: Identification and classification. In M. Pawlak, E. Waniek-Klimczak, and J. Majer (Eds.), *Speaking and instructed foreign language acquisition* (pp. 149-168). Bristol, UK: Multilingual Matters.

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

Cucchiarini, C. & Strik, H. (2018). Automatic speech recognition for second language pronunciation training. In O. Kang, R. I. Thomson, & J. M. Murphy (Eds.), *The Routledge handbook of contemporary English pronunciation* (pp.556-569). New York, NY: Routledge.

Derwing, T., Munro, M., & Carbonaro, M. (2000). Does popular speech recognition software work with ESL speech? *TESOL Quarterly, 34*(3), 592-603.

Fang, T., & Lin, C. (2012). Taiwan EFL learners' pronunciation strategies in two learning contexts. *Journal of Language Teaching and Research, 3*(5), 888-897.

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

Liakin, D. Cardoso, W., & Liakina, N. (2017). Mobilizing instruction in a second-language context: Perceptions of two speech technologies. *Languages*, *2*(3), 1-21.

McCrocklin, S. (2015). Automatic speech recognition: Making it work for your pronunciation class. In J. Levis, R. Mohammed, M. Qian & Z. Zhou (Eds). *Proceedings of the 6th pronunciation in second language learning and teaching conference*. Ames, IA: Iowa State University.

McCrocklin, S. (2016). Pronunciation learner autonomy: The potential of automatic speech recognition. *System, 57*, 25-42.

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

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

McCrocklin, S., Humaidan, A., & Edalatishams, E. (2019). ASR dictation program accuracy: Have current programs improved? In J. Levis, C. Nagle, & E. Todey (Eds.), *Proceedings of the 10th pronunciation in second language learning and teaching conference* (pp. 191-200). Ames, IA, September 2018. Ames, IA: Iowa State University.

Moyer, A. (2014). Exceptional outcomes in L2 phonology: The critical factors of age, motivation, and instruction. *Applied Linguistics, 35*, 418-440.

Mroz, A. (2018). Seeing how people hear you: French learners experiencing intelligibility through automatic speech recognition. *Foreign Language Annals, 51*(3), 1-21.

Osburne, A. (2003). Pronunciation strategies of ESOL learners. *International Review of Applied Linguistics in Language Teaching, 41*(2), 131-143.

Pawlak, M., & Szyszka, M. (2018). Researching pronunciation learning strategies: An overview and a critical look. *Studies in Second Language Learning and Teaching, 8*(2), 293-323.

Peterson, S. S. (2000). Pronunciation learning strategies: A first look (Unpublished research report). ERIC Document Reproduction Service ED 450 599; FL 026 618.

Rabiner, L., & Juang, B. H. (2008). Historical perspective of the field of ASR/NLU. In J. Benesty, M. M. Sondhi, & Y. A. Huang (Eds.), *Springer handbook of speech processing* (pp. 521-538). Berlin, Heidelberg: Springer Berlin Heidelberg.

Strik, H., Neri, A., & Cucchiarini, C. (2008). Speech technology for language tutoring. *Proceedings of language and speech technology (LangTech '08) conference* (pp. 73-76). Rome, Italy.

Wallace, L. (2016). Using Google web speech as a springboard for identifying personal pronunciation problems. *Proceedings of the 7th annual pronunciation in second language learning and teaching conference* (pp. 180-186). Dallas, TX, October 2015. Ames, IA: Iowa State University.

Zielinksi, B., & Yates, L. (2014). Pronunciation is not appropriate for beginning-level students. In L. Grant (Ed), *Pronunciation myths: Applying second language research to classroom teaching* (pp. 56-79). Ann Arbor, MI: University of Michigan Press.