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*, ISSN 2380-9566, Ames, IA, September 2018 (pp. 191-200). Ames, IA: Iowa State University.

**PRESENTATION/POSTER**

# ASR DICTATION PROGRAM ACCURACY: HAVE CURRENT PROGRAMS IMPROVED?

Shannon McCrocklin, Southern Illinois University
Abdulsamad Humaidan, Southern Illinois University
Idée Edalatishams, Iowa State University

Automatic Speech Recognition (ASR) dictation programs have the potential to help language learners get feedback on their pronunciation by providing a written transcript of recognized speech. Early research into dictation programs showed low rates of recognition for non-native speech that prevented usable feedback (Coniam, 1999; Derwing, Munro, & Carbonaro, 2000), but updated research revisiting the accuracy of dictation transcripts for non-native speech is needed. This study investigates current accuracy rates for two programs, *Windows Speech Recognition* (*WSR*) and *Google Voice Typing (Google)*. Participants (10 native English speakers and 20 advanced non-native speakers) read 60 sentences and responded to two open-ended questions. Transcripts were analyzed for accuracy and *t*-tests were used to make comparisons between programs. Major findings include: 1) *Google* displayed a tendency to turn off in the middle of transcription, which affected rates of attempted words; 2) when comparing the accuracy for native versus non-native speech, both programs had higher levels of accuracy for native speech; and 3) when comparing programs for the same speaker, *Google* outperformed *WSR* for both speaker groups on both tasks. Comparing the results to Derwing et al. (2000), *Google* seems to offer substantial increases in accuracy for non-native speakers.

## INTRODUCTION

Automatic Speech Recognition (ASR) is a "machine-based process of decoding and transcribing oral speech" (Levis & Suvorov, 2012, p. 1) that is built into numerous technologies such as automated call centers and dictation programs. ASR technology is also common in language learning software, such as *Rosetta Stone*. ASR has been an interest in the field of pronunciation training since the 1990s partially thanks to the reemerging interest in developing students' spoken language skills (Cucchiarini & Strik, 2018).

Much of the early interest in ASR focused on dictation programs. Dictation programs were developed for native speakers of a given language and are built into both Windows and Mac operating systems as part of their accessibility services. Dictation programs use ASR to interpret what the user has said and provide the spoken utterance in written form. Early tests of the potential of dictation programs for pronunciation practice in a second language highlighted concerns about both the accuracy of the programs for non-native speakers as well as the usability of the feedback (Coniam, 1999; Derwing et al. 2000). Derwing et al. (2000) asked 30 participants (10 native speakers of English, 10 Spanish L1 speakers, and 10 Chinese L1 speakers) to read 60 sentences to *Dragon Naturally Speaking*, a dictation program utilizing ASR. The researchers then assessed the transcription against known intended sentences and against human listeners. They found that while

human listeners were able to understand 95% of the words produced by non-native speakers (and almost 100% of the words produced by native speakers), *Dragon's* accuracy rate was much lower, 71-72% for non-native speakers and 90% for native speakers. Researchers concluded that, given the high levels of inaccurate transcription, use of dictation transcripts would lead to unreliable feedback for second language learners.

In subsequent years, the field instead turned its attention to Computer Assisted Pronunciation Training (CAPT). CAPT programs are developed specifically for non-native speakers of a given language. In a CAPT program, utterances are controlled by having the participant read a written text or respond to a limited range prompt (Hincks, 2015). The program then compares the ASR recognition to the intended response in order to formulate a score or feedback for the student. Within CAPT programs, great strides have been made to improve the accuracy of the evaluation of speech by performing acoustic analysis (Truong, Neri, de Wet, Cucchiarini, & Strik, 2005), incorporating data from non-native speakers (Bouselmi, Fohr, & Illina, 2012; Moustroufas & Digalakis, 2007), examining changes in pronunciation when words are used as part of a larger discourse (Saraçlar, 2000), and hierarchically ranking pronunciation issues based on saliency for more useful feedback (Tepperman, 2009). More importantly, research has shown that CAPT programs can facilitate learning for diverse populations of learners (e.g. children and adults as well as different language backgrounds) (Hincks, 2003; Neri, Cucchiarini, & Strik, 2006; Neri, Mich, Gerosa, & Giuliani, 2008). CAPT programs are, however, limited in their flexibility. Students must follow along prescribed plans of study designed into the CAPT program and must only practice the pre-programmed utterances.

Dictation programs, on the other hand, allow learners to work on whatever topic or language item they wish to. Students could practice words that they struggled with, speak freely to the program on new topics, practice presentations for class, emulate famous speeches, or even read poetry to a dictation program. In recent years, researchers have redeveloped interest in dictation programs for pronunciation practice. Recent research has shown that dictation practice can facilitate student improvement. Research examining practice with dictation programs found that students can improve their production of segmentals using dictation practice with ASR equally well (McCrocklin, 2019) and perhaps even better than when experiencing face-to-face instruction (Liakin, Cardoso, & Liakina, 2014). McCrocklin (2019) focused on a variety of English segmentals (both consonants and vowels) following practice with *Windows Speech Recognition* (*WSR*), while Liakin, Cardoso, & Liaking (2014) focused on the French vowel /y/ following practice with *Dragon Dictation*, a mobile dictation application. However, benefits of ASR practice may extend beyond student improvement. Students reported feeling more empowered in their pronunciation practice when exposed to ASR-based dictation practice (McCrocklin, 2016) and, after using *Google Web Speech (Google)* with pronunciation students, Wallace (2016) argued that ASR dictation practice is useful for getting students to notice pronunciation errors. Wallace described having students dictate while also recording themselves speaking. Students then worked to correct the dictated transcript, using the recording to check what was originally said and to allow for analysis of pronunciation errors. Finally, Mroz (2018) found that students felt that ASR provided a measure of intelligibility useful for understanding how native speakers may perceive their speech.

Despite the recent resurgence of interest, it is unclear to what degree dictation programs have improved in accuracy over the years. McCrocklin (2016) noted that students were still quite frustrated by the number of mistranscribed words provided by *WSR*. To begin answering this question, Edalatishams (2017) compared *Dragon Naturally Speaking* and *Mac Dictation* with 12 sentences (total 58 content words) read aloud by 12 NNSs, finding that Mac Dictation had an average accuracy rate of 77% while Dragon Naturally Speaking had an average accuracy rate of 72%, which was much lower than findings for human listeners (89-98%). These results suggest that perhaps programs have not substantially improved. Without substantial improvement, it is unlikely that programs have moved closer to the goal of providing indications of human intelligibility levels. However, more data is needed and more programs should be tested for accuracy analyses with non-native speakers, ideally using larger data samples.

## THE PRESENT STUDY

The present study examines ASR dictation with specific attention to three dimensions: speakers, dictation program, and speech task. In particular, the preliminary analysis seeks to understand whether changes to speakers, program used, and speech task makes a significant difference to the accuracy of the dictation. Native and non-native speakers of English used *Google* and *WSR* for transcribing read-aloud sentences and free responses.

### Participants

The study included 10 native speakers of American English and 20 non-native speakers whose first languages were Spanish, Chinese, Arabic, French, and Ewe. For this preliminary study, the sample was primarily one of convenience, but efforts were made to gather participants from a variety of language backgrounds. The non-native speakers were advanced English language learners who entered the university with an average TOEFL score of 89.3. The average age of participants, both native and non-native participants was 25.4. The majority of participants were female (60%) in both groups. See Table 1 for more information about each group.

Table 1

*Participant details by language background*

| | Native Speakers | Non-native Speakers |
|---|---|---|
| Number of participants | n=10 | n=20 |
| Average Age | 24 | 26 |
| Gender | Female n=6 | Female n=12 |
| | Male n=4 | Male n=8 |
| Native Language | English n= 10 | Spanish n=7 |
| | | Chinese n=6 |
| | | Arabic n=5 |
| | | French n=1 |
| | | Ewe n=1 |

## Procedures

Participants were provided with information about the study and signed an informed consent form. They were then asked for demographic information through a paper-based questionnaire. For the next stage, participants were recorded as they dictated 60 sentences to two different programs simultaneously, *WSR* and *Google Voice Typing (Google)* in *Drive*, each running on a different computer. To ensure a stable internet connection, which was important for facilitating recognition in *Google*, the computers were hard connected to campus-provided ethernet. Both computers used Logitech USB microphones which were positioned in front of the participant. Participants were also recorded on one computer using *Audacity*. After completing the 60 sentences, participants also responded to two open-ended questions and participated in correcting a copy of the transcript. More information about the dictation and open-response tasks is included in the following sections.

**Dictations.** The participants read aloud sixty sentences to *WSR* and *Google* on the computers, repeating each sentence twice. The second reading provided an opportunity, if needed, for participants to correct a mistake in their reading of the sentence. This step was considered useful for future analyses, although both sentences were included in this pilot and preliminary analysis. The sentences were true/false sentences used successfully in Derwing et al. (2000), adapted for use in this study with the permission of Derwing and Munro. The sentences feature a variety of topics/vocabulary as well as a variety of sounds. The average sentence length was 6.13 words per sentence. For example, a true sentence included was "You can see animals at the zoo."

**Open-ended responses.** The participants also responded to two open ended questions. The dictation programs were turned off during the provision of directions in this stage as well as during the introduction of each open-ended question. Participants were asked to describe how they would plan a surprise birthday party for a friend and to describe their favorite things to do on the weekend. Participants were provided guidance to provide either about 30 seconds of speech or 4-5 sentences in response. After participants had responded to each question, *Audacity* was stopped and both transcribers were turned off. The researcher then made a copy of the transcript provided by one of the dictation programs for both of the open-ended questions. The researcher worked with participants to correct the copied transcripts to provide an accurate transcription of what had been said (to allow for comparisons with the dictated version). During this final step, participants listened to their responses recorded in *Audacity* to remember each utterance and identify mistakes in the transcript.

## Analysis

Analysis included counting the number of words attempted by the program as well as the accuracy of each sentence. Using the accurate list of sentences provided for reading and the accurate, corrected copy of the transcript for the open-ended responses, the ASR-dictated transcripts were scored for the number of accurate words successfully transcribed from the correct version. For this preliminary analysis, the first 10 read sentences, using an average of the first and second attempt, and the first 5 sentences of the open-ended responses were analyzed. Because *Google* had a tendency to turn off, we calculated average number of attempted words as a percentage within each sentence as well as the accuracy of the transcription within the attempted words. For example, if a sentence had 10 words, but *Google* turned off after five, it would have an attempt rate of 50%.

If out of those five, *Google* got four correct, it would have an accuracy rate (within the attempted words) of 80% or 4/5. To compare the accuracy for native speech versus non-native speech on each program (*WSR* and *Google*) for each task (read speech and open-ended response) an independent samples *t*-test was used. Paired samples *t*-tests were used to compare *WSR* and *Google* for 1) each speaker group (native and non-native) and 2) each task (controlled read speech or free open-ended response).

## RESULTS

As mentioned in the analysis section, *Google* had the tendency to turn off frequently during each participant's dictation work. It is not transparent what triggered this issue, but we noticed that hitting enter after each utterance to begin each new sentence on a new line exacerbated the problem. Although we discontinued this practice, *Google* continued to turn off at unpredictable intervals during recoding. Each time it turned it off, we worked to turn it back on at the beginning of a new sentence. The first noticeable finding, then, was that when you include simply a count of accurate transcriptions (thereby counting the non-attempted words against *Google*'s number of accurate transcriptions) *Google* and *WSR* occasionally showed similar numbers of accurate words on certain measures. For example, when examining non-native speakers' read sentences, the accuracy of the words transcribed (when shown as a percentage of all possible read words) was 73.02% for *Google* and 72.1% for *WSR*. In this example, *Google* only attempted 81.04% of the words because of its tendency to shut off, while *WSR* attempted 96.74%. *Google* had a much smaller rate of inaccurate transcriptions at 7.98%, however, while *WSR* showed 24.64% inaccurate transcriptions (see Figure 1).

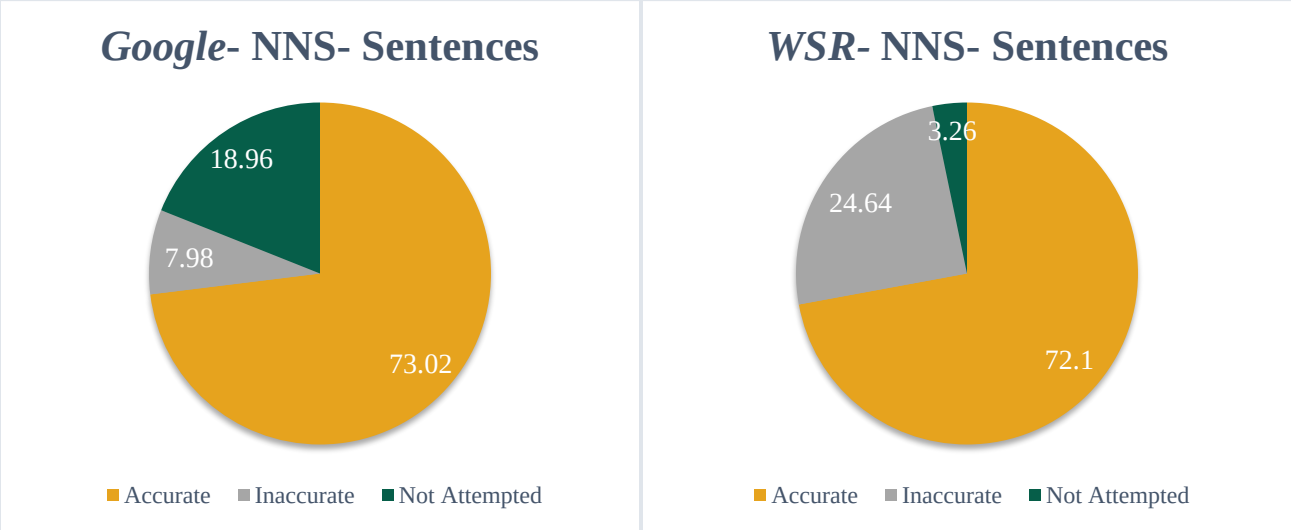


*Figure 1.* Percentage of accurate, inaccurate, and not attempted words for sentences read by non-native speakers by *Google* (left) and *WSR* (right).

Thus, it was important moving forward, to focus on providing the percent of attempted words per sentence as well as the percent of accurate words as a count within the attempted words. Table 2 shows the percentage of attempted words and accuracy among the attempted words for each program on each task by each speaker group (native and non-native).

Table 2

*Mean percentage of attempted and accurate words (within attempted words) by program, task, and speaker group*

| | Sentences | | | | | | Free Speech | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Google* | | | *WSR* | | | *Google* | | | *WSR* | | |
| | Attempted | Accuracy Mean | Accuracy SD | Attempted | Accuracy Mean | Accuracy SD | Attempted | Accuracy Mean | Accuracy SD | Attempted | Accuracy Mean | Accuracy SD |
| Native | 75.74 | 91.95 | 11.25 | 96.72 | 86.81 | 5.14 | 100.00 | 98.00 | 2.58 | 94.50 | 59.70 | 23.41 |
| Non-native | 81.04 | 88.61 | 10.43 | 96.74 | 74.44 | 13.42 | 98.82 | 93.47 | 8.30 | 97.06 | 53.50 | 32.11 |

An independent samples *t*-test was used to compare the accuracy rates for native and non-native speakers on a single program and task. The results showed statistically significant differences between the accuracy rates for *WSR* on the sentences task ($t(28)= 2.79$, $p=.002$) and for *Google* on the free speech task ($t(25)=1.67$, $p<.001$) when comparing native and non-native speakers. Specifically, both *WSR* and *Google* had higher mean accuracy rates for native speakers than for non-native speakers (86.81% vs. 74.4% for *WSR* on sentences and 98% vs. 93.47% for *Google* on free speech), which is in line with previous research from Derwing et al. (2000). In contrast, there were no statistically significant differences when comparing the accuracy of dictation for native versus non-native speakers for *Google* on sentences ($t(28)= .807$, $p=.779$) and *WSR* on free speech ($t(25)= .531$, $p=.470$). This is a surprising finding given that Derwing et al. (2000) previously found significant differences between accuracy rates for native and non-native speakers.

A paired samples *t*-test was used to compare accuracy rates for *Google* and *WSR* for the same speaker population on the two tasks. In all cases, *Google* outperformed *WSR*. The results showed statistically significant differences in three out of the four pairings, non-native speakers on the sentences task ($t(19)=5.42$, $p<.001$) and both native and non-native speakers on the free speech task ($t(9)=5.19$, $p=.001$ and $t(15)=5.14$, $p<.001$ respectively). The differences between *Google* and *WSR* on sentences for native speakers was not statistically significant ($t(9)=1.26$, $p=.238$).

The results further identified an interesting trend: While *Google* became more accurate as speakers switched from the sentence reading task to the free speech, *WSR* displayed an opposite trend. Using a paired *t*-test to compare results from the same speaker on the controlled sentence reading and the free speech task, both trends were statistically significant. For *Google*, in which the system was more accurate in the free speech task, the *p* value was .047 ($t(26)=2.09$), while in *WSR*, in which the system was more accurate in the controlled sentence reading task, the *p* value was less than .001 ($t(26)=-4.78$). Notably, *WSR*'s accuracy on free speech was barely over 50% for non-native speakers. This relationship is illustrated in Figure 2. It is not clear, however, what may have led to such disparities. It is likely that differences in the underlying programming or gaps in the training (voices/styles used to train and check the program during development) have created differential responses to differing speech features.

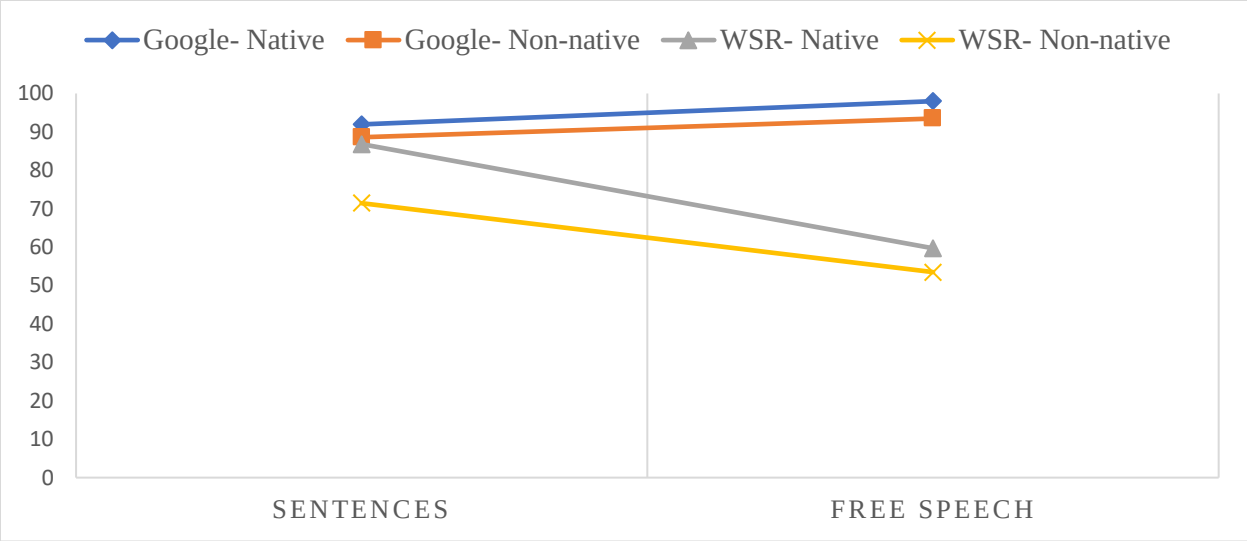


*Figure 2.* Accuracy of attempted words for each program and group of speakers by task type.

## DISCUSSION AND CONCLUSION

The results of this study highlighted several important findings. First, attempted words was an important measure to track. *Google* had the frustrating tendency to turn off which was particularly problematic in the read sentences section, despite changes to protocol to limit the number of stoppages (i.e. forgoing the use of enter to start each sentence on a new line). Notably, although *WSR* outperformed *Google* in attempted words for the sentences task, *Google* outperformed *WSR* in the free speech task. It is not clear what behaviors may have contributed to this difference. Perhaps, because participants were speaking in shorter stretches (usually 4-5 sentences per question) with the dictation programs turned off and restarted for each, *Google* simply had less time to stop working.

Second, despite *Google*'s weakness in turning off, it had much greater accuracy within the attempted words. *Google*'s accuracy for non-native speech ranged from 88.61% (sentences) to 93.47% (free speech), while *WSR*'s accuracy for non-native speakers ranged from 53.50% (free speech) to 74.44% (sentences). *Google*'s outperformance of *WSR* was statistically significant in three out of the four pairings of speaker and task. Comparing our findings to the accuracy rates reported in Derwing et al. (2000), *Google* seems to offer substantial improvement from *Dragon Naturally Speaking* 20 years ago, which offered accuracy rates for non-native speech around 71-72%. *WSR* does not show such improvements.

Further, while the current study did not include comparisons with human raters, *Google* may be getting close to the levels of accuracy of native listeners. While Derwing et al. (2000) found that human listeners were able to transcribe 95-97% percent of non-native speech accurately, Edalatishams (2017) found a range from 89-98% for non-native speech. *Google*'s transcription levels of 88.61% (sentences) to 93.47% (free speech) suggests that *Google* may now rival human listeners particularly for free speech. Further testing is needed and planned for the audio samples in this particular study, however, to make true comparisons to human listeners. Additionally, analysis of speech samples by pronunciation experts in order to examine whether mis-

transcriptions were linked to pronunciation errors will help to determine the usefulness of *Google* for use in second language pronunciation practice. However, an initial examination does show potential for *Google* to be a useful tool. Two examples are illustrated in Table 3 below.

Table 3

*Original sentence, phonetic transcription of utterance, and transcription from Google for 2 participants' utterances*

| | Original Sentence | Phonetic Transcription of Participant Utterance | Sentence Transcription provided by *Google* |
|---|---|---|---|
| Male- Arabic L1 (PSpr18-3) | Most desks are made from spaghetti. | [most dɪsks ɑɹ med fɹəm ˈspaˌgɛti] | Most disks are made from spaghetti. |
| Female- Chinese L1 (PFa17-8) | You can see animals at the zoo. | [ju kæn si ˈæˌnimәs æt di zu] | You can see animals HD 2. |

In this example, although both speakers display multiple pronunciation features that could be labelled as errors, such as the full vowel and stress in the first syllable of "spaghetti" or the mis-stressing and lacking [l] in "animals," *Google* had trouble with "disks" because of the heightened vowel and minimal pair "desks-disks" for the male Arabic speaker. *Google* also had trouble with the full vowels and stressing of "at the" in the female Chinese speaker's utterance which should have been de-stressed as function words. This may additionally indicate that, as *Google*'s ASR has improved, it is less sensitive to accent and more likely to make errors in transcription in places where intelligibility may be negatively impacted for native speakers (for example, in instances of minimal pairs). Having a program that replicated intelligibility for human listeners, as Mroz (2018) has suggested is becoming possible, would be a huge move forward for ASR dictation, making dictation practice more useful for second language learners. Further analysis and testing is needed.

## ACKNOWLEDGEMENTS

We would like to thank Tracy Derwing and Murray Munro for providing the sentences originally used in Derwing et al. (2000) and for allowing their adaptation into this study.

## ABOUT THE AUTHORS

Shannon McCrocklin (shannon.mccrocklin@siu.edu) is an Assistant Professor at Southern Illinois University in the Department of Linguistics. Her primary research area focuses on second language pronunciation teaching methods and, specifically, ways of incorporating technology into second language learning.

Contact Information:
1000 Faner Dr., Rm 3228
Southern Illinois University
Carbondale, IL 62901
1-618-453-3428

Abdulsamad Humaidan (humaidanabdulsamad@siu.edu) is a doctoral candidate at Southern Illinois University pursuing his PhD in Education focusing on Curriculum and Instruction and TESOL. His research interests include ESL and EFL assessment, teacher education, and language teaching and learning with technology.

Idée Edalatishams (edalati@iastate.edu) is a graduate student at Iowa State University pursuing her PhD in Applied Linguistics and Technology. Her research is in the fields of sociolinguistics and discourse analysis using corpus methodology with a primary focus on discourse intonation and non-native speakers' identity.

## REFERENCES

Bouselmi, G., Fohr, D., & Illina, I. (2012). Multilingual recognition of non-native speech using acoustic model transformation and pronunciation modeling. *International Journal of Speech Technology*, *15*, 203-213.

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.

Edalatishams, I. (Sept, 2017). *Non-native speech and recognition accuracy of two ASR applications: Dragon and Dictation*. Paper presented at the meeting of Pronunciation in Second Language Learning and Teaching, Salt Lake City, UT.

Hincks, R. (2015). Technology and leaning pronunciation. In M. Reed & J. Levis (Eds.), *The handbook of English pronunciation* (pp. 505-519). Malden, MA: John Wiley & Sons, Inc.

Hincks, R. (2003). Speech technologies for pronunciation feedback and evaluation. *ReCALL*. *15*(1), 3-20.

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

Levis, J. & Suvorov, R. (2012). Automated speech recognition. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Retrieved from http://onlinelibrary.wiley.com/

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

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

Moustroufas, N. & Digalakis, V. (2007). Automatic pronunciation evaluation of foreign speakers using unknown text. *Computer Speech and Language, 21*, 219-230.

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

Neri, A., Cucchiarini, C., & Strik H. (2006). ASR-based corrective feedback on pronunciation: does it really work? *Proceedings of the ISCA Interspeech 2006*, Pittsburgh, PA, 1982–1985.

Neri, A., Mich, O., Gerosa, M., & Giuliani, D. (2008). The effectiveness of computer assisted pronunciation training for foreign language learning by children. *Computer Assisted Language Learning, 21*(5), 393-408.

Saraçlar, M. (2000). *Pronunciation modeling for conversational speech recognition* (Doctoral dissertation). Ann Arbor, MI: UMI Dissertation Services.

Tepperman, J. (2009). *Hierarchical methods in automatic pronunciation evaluation.* (Doctoral dissertation). Ann Arbor, MI: UMI Dissertation Services.

Truong, K., Neri, A., de Wet, F., Cucchiarini, C, & Strik, H. (2005). Automatic detection of frequent pronunciation errors made by L2-learners. *Proceedings from InterSpeech 2005 (IS2005),* Lisbon, Portugal, 1345-1348.

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*. Retrieved from https://apling.engl.iastate.edu/alt-content/uploads/2016/08/PSLLT7_July29_2016_B.pdf.