QUÃO BEM A TECNOLOGIA RAF PODE ENTENDER A FALA COM SOTAQUE ESTRANGEIRO?

Hanna Kivistö Souza, William Gottardi

https://doi.org/10.1590/010318138668782v61n32022

Submetido em: 2022-10-26
Postado em: 2022-10-26 (versão 1)
(AAAA-MM-DD)

Trabalhos em Linguística Aplicada renders public this preprint hosted on SciELO Preprints, which is an integral part of the SciELO collection. The text was formally approved after peer review. Its final version, with possible modifications, will be published in an upcoming issue of the journal.

HOW WELL CAN ASR TECHNOLOGY UNDERSTAND FOREIGN-ACCENTED SPEECH?

QUÃO BEM A TECNOLOGIA RAF PODE ENTENDER A FALA COM SOTAQUE ESTRANGEIRO?

Hanna Kivistö de Souza[*, **]
William Gottardi[***, ****]

ABSTRACT
Following the Covid-19 pandemic, digital technology is more present in classrooms than ever. Automatic Speech Recognition (ASR) offers interesting possibilities for language learners to produce more output in a foreign language (FL). ASR is especially suited for autonomous pronunciation learning when used as a dictation tool that transcribes the learner's speech (McCROCKLIN, 2016). However, ASR tools are trained with monolingual native speakers in mind, not reflecting the global reality of English speakers. Consequently, the present study examined how well two ASR-based dictation tools understand foreign-accented speech, and which FL speech features cause intelligibility breakdowns. English speech samples of 15 Brazilian Portuguese and 15 Spanish speakers were obtained from an online database (WEINBERGER, 2015) and submitted to two ASR dictation tools: Microsoft Word and VoiceNotebook. The resulting transcriptions were manually inspected, coded and categorized. The results show that overall intelligibility was high for both tools. However, many features of normal FL speech, such as vowel and consonant substitution, caused the ASR dictation tools to misinterpret the message leading to communication breakdowns. The results are discussed from a pedagogical viewpoint.
Keywords: intelligibility; automatic speech recognition; L2 pronunciation development; autonomous learning.

RESUMO
Após a pandemia de Covid-19, as tecnologias digitais estão mais presente nas salas de aula do que nunca. O Reconhecimento Automático da Fala (RAF) oferece possibilidades interessantes para os aprendizes de uma língua estrangeira (LE) aumentarem sua produção oral. O RAF é especialmente adequado para a aprendizagem autônoma de pronúncia quando usado como uma ferramenta de ditado que transcreve a fala do estudante (McCROCKLIN, 2016). No entanto, as ferramentas de RAF são treinadas com falantes nativos monolíngues em mente, não refletindo a realidade dos falantes de inglês em uma escala global. Consequentemente, o presente estudo examinou quão bem duas ferramentas de ditado que utilizam ASR entendem a fala com sotaque estrangeiro e quais características causam falhas de inteligibilidade. Amostras de fala em inglês de 15 falantes de português brasileiro e 15 falantes de espanhol foram obtidas de um banco de dados online (WEINBERGER, 2015) e submetidas a duas ferramentas de ASR: Microsoft Word e VoiceNotebook. As transcrições foram manualmente inspecionadas, codificadas e categorizadas. Os resultados mostram que a inteligibilidade geral dos falantes foi alta para ambas as ferramentas. No entanto, muitas características normais, como modificações vocálicas e consonantais, da fala em LE fizeram com que as ferramentas de ditado ASR interpretassem mal a mensagem, levando a falhas de comunicação. Os resultados são discutidos do ponto de vista pedagógico.
Palavras-chave: inteligibilidade; reconhecimento automático da fala; desenvolvimento de pronúncia em LE; aprendizagem autônoma.

INTRODUCTION

Acquiring an intelligible pronunciation in a second language (L2) can be an arduous task, which many L2 learners face mainly alone, as pronunciation teaching notoriously receives little time in foreign language classrooms (DERWING, 2010). In order to complement their in-class pronunciation learning, learners can employ

* Universidade Federal de Santa Catarina (UFSC), Florianópolis, SC, Brasil.
** University of Turku, Turku, Finland. hanna.kivistodesouza@gmail.com
Orcid: https://orcid.org/0000-0002-8498-2691
*** Universidade Federal de Santa Catarina (UFSC), Florianópolis, SC, Brasil. , teacher.will@outlook.com
**** This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.
Orcid: https://orcid.org/0000-0002-1291-3953

autonomous learning activities on the road to a more intelligible pronunciation (CARLET; KIVISTÖ DE SOUZA, 2018). Speech technology such as pronunciation apps, websites, computer programs, and dictation tools offers interesting possibilities for this end. From the growing body of research on the field of L2 pronunciation and speech technology (ASHWELL; ELAM, 2017; INCEOGLU; LIM; CHEN, 2020; LIAKIN; CARDOSO; LIAKINA, 2017; McCROCKLIN; EDALATISHAMS, 2020; MROZ, 2018), it is clear that the topic has called researchers' interest, becoming an area of its own – computer-assisted pronunciation teaching (CAPT).

L2 speech research has witnessed two important changes of foci in the last decades. For one, the focus has moved from aiming for native-like speech to aiming for intelligible speech (LEVIS, 2005). On the other hand, the context of interaction has expanded from the traditional native speaker - non-native speaker interaction to embrace interaction between non-native speakers of English (JENKINS, 2002; JENKINS; COGO; DEWEY, 2011). Jenkins' English as an International Language (EIL) approach seems especially suited for the Latin American context, where English learners carry out most of their interaction with other non-native speakers of English.

Following Munro and Derwing (1995), we understand *intelligibility* as the degree to which the speaker's message is actually understood by the listener. Intelligibility is independent, but somewhat related to *accentedness* (the degree to which one's speech diverges from the local variety) and *comprehensibility* (the amount of effort the listener puts in to understand the message) (DERWING; MUNRO, 1997). In other words, whereas accentedness and comprehensibility deal with the listeners' perception of the speech, intelligibility is involved with the actual comprehension of the message. Consequently, accentedness and comprehensibility are frequently assessed in Likert scales whereas intelligibility is frequently assessed through transcribing what was actually heard (MUNRO; DERWING, 1995).

Assessing intelligibility through the transcription of learners' utterances can be carried out through Automatic Speech Recognition (ASR) technology. ASR is "an independent, machine-based process of decoding and transcribing oral speech" (LEVIS; SUVOROV, 2013, p. 316). In addition, ASR can be used in human-computer interaction through auto-generated captions, dictation tools, and Intelligent Personal Assistants (IPAs) such as Microsoft Cortana, Google Assistant, Siri, and Alexa (INCEOGLU; LIM; CHEN, 2020).

ASR programs can be helpful in L2 pronunciation development as they can increase exposure to the language outside the classroom (CARLET; KIVISTÖ DE SOUZA, 2018; LIAKIN; CARDOSO; LIAKINA, 2017), help motivate the language learners (LEVIS; SUVOROV, 2013; MROZ, 2018), contribute to autonomous learning (INCEOGLU; LIM; CHEN, 2020; McCROCKLIN, 2016), and offer additional possibilities for output (DIZON; TANG, 2020; LIAKIN; CARDOSO; LIAKINA, 2017). Furthermore, ASR-based dictation tools may aid the learner to notice the gap between their output and target-like production. Noticing the gap could take place in situations in which the learner speaks to the program and its response (transcription or action, such as "play the song X") does not match the intended message. This would be the case of a Brazilian learner of English asking Microsoft Cortana the meaning of the word "event", but misplacing the stressed syllable as ['i.vent], and the smart assistant answering the definition of the word *even* (['i.vən]) instead. Another example would be to use an online dictation tool and to read aloud a passage by comparing the output text with the intended one. This last affordance of ASR technology is the focus of this paper.

This study examines how well ASR-based dictation tools that are freely available for language learners can understand foreign-accented speech, and consequently, discusses whether such programs could be useful for L2 pronunciation development, especially in contexts where there is little exposure to the foreign language. In the following section, we revise some concepts about ASR technology and discuss previous research on ASR and L2 learning with focus on L2 speech acquisition. We will then present the methodology of the present study and the results. Finally, the findings are discussed from the point of view of L2 pronunciation teaching and autonomous learning.

1. ADVANCES AND LIMITATIONS OF ASR

ASR systems transcribe oral input into word sequences, which are their output (YU; DENG, 2015). The first ASR systems were developed in the early 1950s (LEVIS; SUVOROV, 2013) and, throughout these decades much progress has been made due to new model techniques, more robust algorithms, enhancement of noisy speech recognition, and the growing demand for its integration with mobile and smart devices (JURAFSKY; MARTIN,

2021; LEVIS; SUVOROV, 2013). Moreover, the ASR system's accuracy rates are constantly increased by using machine learning with relevant acoustic information (ASHWELL; ELAM, 2017).

ASR systems share some characteristics which pose limitations. If the vocabulary present in the input string sent to the system is large, the task will be much harder. A system that expects a *yes*/*no* answer as the input string will transcribe much more accurately than a transcription of a lecture, for instance. Due to predictability, Human-Machine Communication is easier to recognize than Human-Human Communication once conversational speech predicts more lexical variation and a higher speech rate. Channel and noise also impact automatic speech recognition. ASR systems work more accurately in quiet environments with proper input devices, such as head-mounted microphones. Finally, speaker characteristics affect the working of ASR programs: Speech recognition is easier if the speaker is speaking in the same variety that the ASR system was trained on. Moreover, recognition of child speech is harder if the system was trained only on adult speech (JURAFSKY; MARTIN, 2021).

Considering these dimensions, some limitations can be predicted for ASR systems. They still perform poorly under specific conditions despite the recent technological advancements. These conditions include far-field microphones, very noisy environments, accented speech, side talks, and multitalker speech, and spontaneous speech which may be disfluent, with emotion, or with variable speed (YU; DENG, 2015). Furthermore, the complexity of natural languages and their lexical, semantic, and phonological variation poses challenges for ASR. ASR systems need to be able to recognize units (e.g., phones, syllables, words, and phrases) accurately, and make coherent lexical choices despite syntactic and semantic ambiguity (LEVIS; SUVOROV, 2013). Therefore, ASR systems can be evaluated through a standard evaluation metric called Word Error Rate (WER). WER is based on how much the transcription differs from the intended message: thus, the lower the WER, the more accurate the ASR system is. WER can vary from the range of 20% in ideal conditions to as high as 81% in a multispeaker environment (JURAFSKY; MARTIN, 2021; YU; DENG, 2015). Notwithstanding, WER can be much lower in controlled environments. For instance, for tasks such as voice search and mobile short message, WER is considerably below 10 % (YU; DENG, 2015).

Although ASR technology faces some challenges, much improvement has been achieved, which has allowed ASR to play an important role in many applications. Some of these applications help to improve communication between people (e.g., speech-to-speech translation systems, dictating short messages, and automatic captioning). Other applications facilitate communication between humans and machines (e.g., gaming, personal digital assistants, and voice search) (JURAFSKY; MARTIN, 2021; YU; DENG, 2015). The next section looks at the role ASR technology may play for L2 learning.

2. ASR TECHNOLOGY AND L2 LEARNING

ASR can also be used for educational purposes. Levis and Suvorov (2013) suggest that ASR can be useful for teaching children to read, providing relevant feedback during speaking activities, and allowing learners to produce more speech. In a review of over 350 studies, Golonka et al. (2014) found strong support for the effectiveness of ASR tools on foreign language learning. The authors focused on empirical studies that compared the use of traditional materials and methods to newer technologies. The review included individual study tools (e.g., electronic dictionary, grammar checker, ASR, and pronunciation programs), mobile devices, classroom-based technologies, and network-based social computing. For them, ASR can make a considerable impact in foreign language learning, mainly because of its potential to provide learners with feedback and help them improve their pronunciation. The authors also found that learners tend to have positive attitudes towards the learning process when using digital technology in general, being more motivated to use it instead of traditional materials.

ASR thus becomes a versatile resource for teaching and autonomous learning mainly because it is now available as a built-in feature in many different free programs on the internet (e.g., Google Docs' voice typing, Google Translator, or Bing and Google's voice web search) (YOSHIDA, 2018). This vast number of ASR programs allows the language learner to speak the foreign language and see it converted instantaneously into written words. The current state of ASR technology can make this conversion much more accurate if compared to a decade ago and some ASR programs can transcribe texts into over one hundred languages (HENRICHSEN, 2020).

Out of the four skills, the learning of L2 speech might benefit from ASR technology the most. The objective of using ASR with this aim is to aid learners to have a more intelligible pronunciation (YOSHIDA, 2018). In the

last decades, studies involving ASR programs point to benefits of this technology for pronunciation learning. ASR programs tend to be user-friendly, increase learners' motivation, and foster learner autonomy (KIM, 2006; LIAKIN; CARDOSO; LIAKINA, 2015, 2017; McCROCKLIN, 2014, 2016; MROZ, 2018). For example, Mroz (2018) found that French learners using ASR in Gmail as a way of creating awareness of their own intelligibility described their experience as having a positive influence on their learning process. Similarly, Kim (2006) presents that ASR-based pronunciation training can create a safe environment where learners feel more confident during their practices. Moreover, Liakin et al. (2017) reports the results of empirical studies investigating the usage of ASR and text-to-speech to promote L2 French pronunciation improvement. They found that students enjoy using this technology especially because of the mobile-enhanced learning environment that those technologies provide learners with, besides fostering their autonomy. In addition, McCrocklin (2016) shows that, after a 3-week pronunciation workshop, students revealed that feedback provided by the ASR program allowed them to practice autonomously increasing thus their beliefs of autonomy.

Notwithstanding, ASR can also bolster learners' confidence and willingness to communicate, and reduce foreign language anxiety (BASHORI et al., 2020, 2021; CHEN, 2011; KIM, 2006; MROZ, 2018). Furthermore, ASR programs are especially useful in a foreign language context in which learners have limited contact with other L2 speakers; hence, fewer opportunities to practice their pronunciation skills (DIZON; TANG, 2020). The feedback provided by ASR has been shown to facilitate the improvement of problematic speech sounds, mainly on the segmental level (NERI; CUCCHIARINI; STRIK, 2006, 2008). The authors further state that pronunciation improvement was achieved after a few hours over one month of practice using an ASR system developed by them (CUCCHIARINI; NERI; STRIK, 2009). Nevertheless, it is important to bear in mind that ASR-based programs should be used as complements to in-class pronunciation instruction, not as substitutes for teacher-led pronunciation instruction (CUCCHIARINI; STRIK, 2018). Finally, ASR's characteristics make it a versatile tool to be used easily outside the classroom (LIAKIN; CARDOSO; LIAKINA, 2015, 2017).

Consequently, ASR-based dictation tools are a viable option for extended pronunciation practice. They are more accessible and flexible to use if compared to other computer-assisted pronunciation training programs (McCROCKLIN, 2019a), which makes them excellent for pronunciation learning. Furthermore, dictation programs allow different types of practice from single words to whole sentences of varied topics of interest. Nonetheless, despite these advantages, dictation programs have not been enough explored due to their low levels of speech recognition, particularly considering nonnative speakers (McCROCKLIN, 2019b). As McCrocklin and Edalatishams (2020) point out, ASR technology recognized nonnative speech with 18-20% less accuracy than native speech in the past; however, recent improvements in ASR technology have allowed Google Voice Typing to reduce this transcription accuracy gap to only 3-5%. Few studies exist on the use of ASR in L2 pronunciation development, and it is necessary to investigate the accuracy rates of different dictation programs for nonnative speech (McCROCKLIN, 2019a). This is the objective of this paper as the next section presents.

3. RESEARCH QUESTIONS AND OBJECTIVES

The objective of this paper was to study the effectiveness of two ASR-based dictation programs (Microsoft Word dictation and VoiceNotebook) in assessing the intelligibility of Latin American English users' speech. Our aim was to examine whether the two dictation programs in question showed any differences in assessing intelligibility and what pronunciation deviations caused intelligibility breakdowns. Furthermore, by examining two L1 groups, we sought to extend the amount of possible pronunciation deviations the dictation programs would encounter. The following questions were posed:

RQ1: Do Microsoft Word dictation and VoiceNotebook judge intelligibility differently and if so, to which ASR-based dictation program, the EIL speakers are more intelligible?
RQ2: Are the EIL speakers from L1 Spanish and L1 Portuguese equally intelligible to the ASR-based dictation programs?
RQ3: What compensatory strategies do the ASR-based dictation programs use in cases of intelligibility breakdowns?
RQ4: Which pronunciation deviations cause intelligibility breakdowns for the ASR-based dictation programs?

4. METHOD

The speech samples to this study came from the Speech Accent Archive (WEINBERGER, 2015). The Speech Accent Archive is an online free-access database that contains paragraph reading recordings in English by English language users. The elicited paragraph was designed to be phonetically variable, representing all the English segments in varied phonetic environments. The paragraph the speakers read is as follows:

> Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station. (WEINBERGER, 2015)

From the Speech Accent Archive, a selection of the speakers was made based on the following criteria: birthplace, native language, age of onset of learning (AOL), and length of residence (LOR). As our intention was to assess the intelligibility of EIL users, we wanted to include speakers whose LOR would be as short as possible and also speakers who had started studying English well after having acquired their L1. We expected that by selecting speakers with these characteristics, the speech samples would contain pronunciation deviations from General American English (GA). Speakers whose LOR was longer than 1.5 years were disregarded. To assess the baseline performance of the two ASR-based dictation programs, and to assure that the programs were suitable for the purpose of the study, we selected five native English speakers. These speakers were randomly chosen and came from the states of Iowa, California, Massachusetts, Ohio, and California.

As for the EIL groups, Brazilian Portuguese (BP) and Spanish (Sp) were chosen, as these two EIL groups bear special importance in the Latin American setting and make up important immigrant populations in the United States. All the Brazilian Portuguese speakers had been born in Brazil and reported Portuguese as their native language. Our objective was to obtain L1 Spanish speakers from the same country in order to avoid dialectal differences in the L1 that could cause differences in the L2 phonological learning[1]. We chose Venezuela for its geographical proximity to Brazil and the great number of Venezuelan immigrants in the US. However, after applying the LOR and AOL criteria, there were not enough speakers from Venezuela. Consequently, we also included speakers from Colombia, expecting that the dialectal differences between Venezuelan and Colombian Spanish would not be enough to result in L2 phonological learning differences. Six of the L1 Sp speakers were born in Venezuela and nine in Colombia.

Speakers' demographic variables can be observed in Table 1. *Age of Onset of Learning* was obtained from the answer to the question: "How old were you when you first began to study English?". In other words, it did not discriminate between naturalistic and academic settings. *Length of Residence* was understood as the amount of time the subjects had lived in an English-speaking country. Based on subjects' age and AOL, we calculated a measure of L2 experience (Age-AOL). The variable *Number of foreign languages* was obtained by adding up the languages other than the L1 the subjects reported to know. A series of independent samples t-tests showed that the two EIL groups did not differ significantly in their AOL ($t(20,5) = -1.68$, $p = .10$) , LOR ($t(28) = -.56$, $p = .57$) or L2 Experience ($t(22,3) = -.025$, $p = .98$).

Table 1. Speaker characteristics

Group	Age	AOL	LOR	L2 Experience	Nº of FLs
L1 Eng ($n=5$)	23 (6.89) 18-35	-	-	-	0.40 (0.54) 0-1
L1 BP ($n=15$)	22.40 (3.66) 18-29	10.33(2.32) 8-15	0.37 (0.48) 0-1.2	12.07 (5.21) 3-20	2.20 (1.14) 1-5
L1 Sp ($n=15$)	24.73 (5.67) 17-34	12.60 (4.65) 4-20	0.47 (0.49) 0-1.5	12.13(9.04) 1-27	1.19 (1.33) 1-6

Source: authors

Age, AOL, and LOR expressed in years. Standard deviation between brackets. Range of responses on the second line.

1. For example, in Castilian Spanish, spoken in the north and central parts of Spain, the phone [θ] is a phoneme whereas in Latin America this realization is not present having been substituted with /s/. As English has /θ/ (e.g., 'thing'), the status of the L2 sound (old/new) is different for the speakers of different varieties of Spanish.

5.1 Transcription instruments and procedures

Two ASR-based dictation programs were used to transcribe the samples. The programs were selected under the criteria that they must have a free version and they are available on different platforms. This decision was made because we wanted to analyze tools that are more easily accessible for the largest number of language learners.

The first dictation tool is embedded into Microsoft Word[2] (from now on referred to as MW dictation). It is available for free on its web version, which can be accessed on different platforms using the web browsers Edge, Firefox, or Chrome (MICROSOFT, 2021). The second dictation tool is the website VoiceNotebook[3]. As the former tool, VoiceNotebook is also a voice recognition application that converts speech to text. It uses Google Chrome's implementation of Web Speech Application Programming Interface[4] (API) as informed by the website's developer. The same API can be used for different websites and extensions for speech synthesis and recognition if accessed through Google Chrome web browser. The website is free and works in multiple operating systems (Windows, Mac, and Linux OS). Furthermore, it is available as a mobile application and a Google Chrome extension that allows users to utilize voice input on text fields of any website (VOICENOTEBOOK, 2021).

To obtain the transcriptions for the intelligibility analysis, the following procedures were followed. Each speech sample was first downloaded from the Speech Accent Archive database as an audio file to a PC. Next, the audio files were played individually using an in-ear headphone. The headphone's speaker worked as the output device, and the headphone's microphone worked as the input device; that is, the audio was played from the speaker and captured by the microphone on the same computer. To reduce outside noise, the headphone was placed on a cardboard box. Then, a single speaker of the headphone was put right beside the headphone's microphone. Each audio file was played without pausing until the dictation program had finished transcribing the sample. The same procedure was followed for both dictation programs and each speech sample. Both programs were configured to use American English voice recognition. Moreover, all audio files were played in quiet moments of the day under the same conditions. Whenever a distinctive background noise was detected, the transcription was restarted.[5] After transcribing all the samples, the data was coded and analyzed as seen in the next section.

5.2 Analyses

The transcription data were manually inspected and coded by one of the researchers. The errors in the transcripts, or as we understand them, the compensatory strategies employed by the programs in case of intelligibility breakdowns, were marked. From the inspection of the data, six compensatory strategy types were observed, and the errors committed by the programs were subsequently classified into one of these categories: substitution, omission, addition, 2x1 substitution, and 1x2 substitution. *Substitution* involved the substitution of the original word with a different one (e.g., *forms* instead of *spoons*). *Omission* involved the omission of a word that was present in the recording (e.g., *small plastic snake* instead of *a small plastic snake*). *Addition* indicated the addition of a word not present in the recording (e.g., *scoop of these things* instead of *scoop these things*). *2x1 substitution* involved the substitution of two words with a single one (e.g., *chicken* for *she can*). Finally, *1x2 substitution* indicated that the program had substituted one word from the recording with two new ones (e.g., *and no* instead of *snow*). In the data coding, homophones were not considered errors (e.g., *read-red* or *peas-Ps*). Singular forms instead of plural (e.g., *spoon* instead of *spoons*) or genitive instead of plural (*thing's* instead of *things*) were also not considered errors.

A mixed-methods approach was adopted. An intelligibility score was calculated for each participant in both programs by dividing the number of correctly transcribed words by the total number of words and by multiplying it by 100. The total number of words in the Speech Accent Archive paragraph is 69. Some speakers omitted words from the recording, and in those cases, the actual number of words uttered by the speaker was used. Additionally, we calculated the total number of compensatory strategies by each speaker for both programs and the number of compensatory strategies by their type. For the qualitative analysis, we analyzed the data for the substitution patterns

2. Desktop version 2104, from May 11th, 2021.
3. https://voicenotebook.com/
4. https://wicg.github.io/speech-api/
5. In the case of six speakers, the original sound quality was non-optimal. The sound quality was improved with noise-reduction treatment. However, as the treatment did not have an effect on the resulting transcripts, the noise-reduced sound files were disregarded, and the original ones were used in all analyses.

and noted down which were the items that caused more intelligibility breakdowns, and which were the most frequent lexical substitutions employed by the programs.

5. RESULTS

In this section we look at the performance of the two ASR-based dictation programs in regard to the overall intelligibility, the use of compensatory strategies, and the items and pronunciation deviations that caused intelligibility breakdowns. Normality analyses indicated that the intelligibility score data were normally distributed, but the compensatory strategies data were skewed leading to the use of parametric tests for the first case and non-parametric tests for the latter case.

Descriptive statistics of the overall intelligibility data can be observed in Table 2 below. Comparing the intelligibility of the native English speakers with the EIL speakers, it can be observed that whereas the native speaker intelligibility was higher for both programs than the mean intelligibility of the two EIL groups, there was still variation in the native speaker intelligibility. In fact, one of the native speakers, *eng200*, obtained an intelligibility score of 85.5% which falls within the intelligibility range of the L1 BP and L1 SP speakers. An auditory inspection of *eng200* speech sample revealed no audible pronunciation deviations but a high speech rate. We can thus hypothesize that not only pronunciation deviations, but also other characteristics of speech, such as speech rate, affect the intelligibility for ASR-based dictation programs. From the programs under analysis, especially VoiceNotebook appeared to be negatively impacted by high speech rate.

Table 2. Mean number of intelligibility breakdowns per speaker and mean intelligibility score

Groups	Number of intelligibility breakdowns averaged across speakers		Intelligibility Score (%)	
	VN	MW	VN	MW
L1 Eng (n=5)	3.20 (3.89) 1-10	0.80 (0.83) 0-2	95.36 (5.64) 85-98	98.33 (1.21) 97-100
L1 BP (n=15)	11.27 (5.71) 5-27	4.93 (2.93) 1-11	83.59 (8.36) 60-92	92.80 (4.30) 83-98
L1 Sp (n=15)	16 (10.32) 3-38	10.20 (8.31) 0-34	76.76 (15.06) 44-95	85.19 (12.06) 50-100

Source: authors

Standard deviations between brackets. Range on the last line rounded to whole numbers. VN= Voice Notebook. MW = Microsoft Word dictation

In order to determine whether the two programs judged intelligibility of the EIL speakers differently (RQ1), a paired samples t-test was conducted to compare the mean intelligibility scores of the EIL speakers between VoiceNotebook (M= 80.18, SD= 12.46, n=30) and MW dictation (M= 89.00, SD= 9.70, n= 30). The results showed a significant difference in the means, indicating that the EIL speakers as a group were more intelligible to MW dictation than to VoiceNotebook ($t(29)$=-6.07, $p<.001$). The eta squared (.55) indicated a large effect size.

Research question 2 asked whether the two EIL speaker groups (L1 Brazilian Portuguese and L1 Spanish) were equally intelligible to the ASR-based dictation programs. Two independent samples t-tests were conducted to compare the intelligibility scores of the L1 BP and L1 Sp speakers between the two programs. The results indicated that the intelligibility of the L1 BP speakers (M=83.59, SD=8.36) was not significantly different from the L1 Sp speakers (M= 76.76, SD= 15.06) for VoiceNotebook ($t(28)$= 1.53, p).13. For MW dictation, the L1 BP speakers (M= 92.80, SD= 4.30) were significantly more intelligible than the L1 Sp speakers (M= 85,19, SD= 12.06, $t(17,5)$= 2.30, $p<.03$).

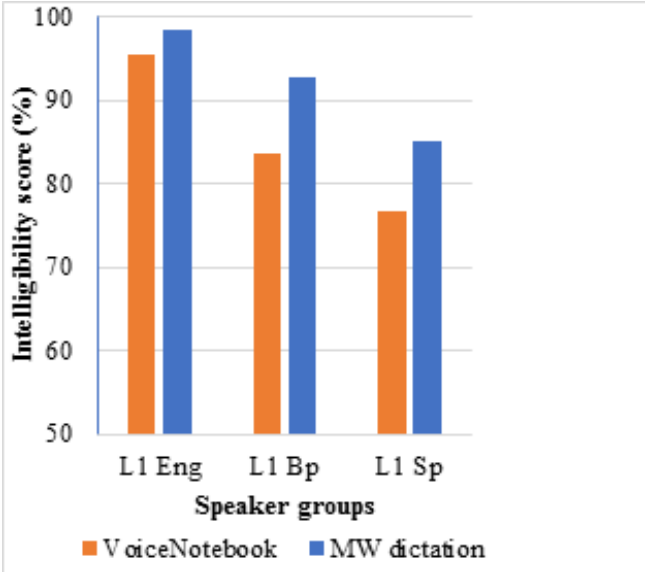

Figure 1. Mean intelligibility of the L1 English speakers and the two EIL groups across the programs.
Source: authors

After determining the overall intelligibility of the speakers, the intelligibility breakdowns were examined. In order to determine what compensatory strategies VoiceNotebook and MW dictation used when the transcription differed from the intended message (RQ3), the amount of times each compensatory strategy was employed was inspected. Table 3 below indicates the mean amount of times a given compensatory strategy was used by the two programs per IEL speaker.

Table 3. Number of compensatory strategies averaged across the L2 speakers by the two programs

***n*=30**	Voice Notebook	MW dictation
Substitution	8.37 (5.16)	5.40 (4.91)
Omission	2.97 (6.21)	0.50 (0.82)
Addition	0.57 (0.89)	0.07 (0.36)
2x1 Substitution	0.73 (0.64)	0.70 (0.95)
1x2 Substitution	0.27 (0.69)	0.20 (0.55)

Source: authors
Standard deviations between brackets

As can be observed, both programs employed *Substitution* more than other strategies. On average, VoiceNotebook substituted 8.37 words per speaker, whereas MW dictation performed 5.4 substitutions per transcription. In order to determine whether there were statistical differences in the use of the compensatory strategies, a Friedman test was carried out separately for MW dictation and VoiceNotebook with Compensatory Strategy Type (*Substitution/ Omission/ Addition/ 2x1 Substitution/ 1x2 Substitution*) as the independent variable and *Compensatory Strategy* as the dependent variable.[6]

For VoiceNotebook, Friedman test showed that there was a significant difference in Compensatory Strategy Type ($\chi 2$ (4)= 70.90, $p<.001$). Post-hoc Bonferroni adjusted Wilcoxon signed-rank tests indicated that *Substitution* was significantly different from the other types ($p<.001$). *Omission* was significantly higher than *2x1 Substitution* ($Z=-2,91$, $p=.004$) and *1x2 Substitution* ($Z=-2.95$, $p=.003$). The remaining comparisons were not statistically significant ($p>.05$).

For MW dictation, The Friedman test also revealed a significant difference in the Compensatory Strategy Type ($\chi 2$ (4) =77.51, $p<.001$). Post-hoc Bonferroni adjusted Wilcoxon signed-rank tests further showed that as for VoiceNotebook, *Substitution* significantly differed from the other types ($p<.001$). *Addition* and *2x1 Substitution*

6. A set of Mann Whitney U-tests was conducted to compare the compensatory strategy types between the two L1 groups. As no significant differences were found, the L1 BP and L1 Sp speakers were grouped together for this analysis.

also differed significantly from each other (Z=-2.80, p=.005). The remaining comparisons were not statistically significant.

Finally, in order to compare whether the frequency of use of a specific compensatory strategy differed between the programs, a series of Wilcoxon Signed Rank tests was carried out. The difference in the strategy use between the two programs was statistically significant in the case of *Substitution* (Z=-3.71, p<.001), *Omission* (Z=-2.62, p=.009), and *Addition* (Z=-2.41, p=.016), in the way that VoiceNotebook employed these strategies significantly more than MW dictation. There was no statistical difference in the frequency of use of *2x1 Substitution* or *1x2 Substitution* between the two programs.

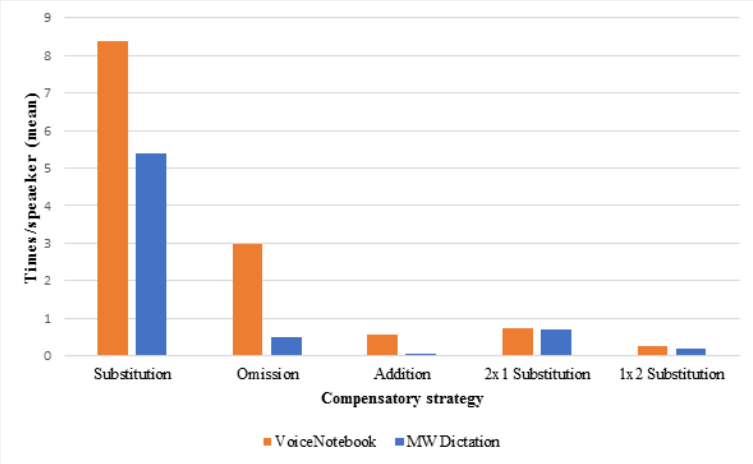

Figure 2. Use of compensatory strategies by the two programs averaged across speakers
Source: authors

Observing the data qualitatively, words that caused intelligibility breakdowns for the ASR programs were identified (RQ4). Due to space constraints, a cut-off line at 10 wrongly-transcribed words was established for reporting. Table 4 below indicates the lexical items that caused ten or more intelligibility breakdowns, considering the Brazilian Portuguese, Spanish and American English speakers together (n=35). The words *slabs*, *thick* and *spoons* were especially problematic for the two dictation programs. It can be observed that, except for the conjunction *and*, all the items that resulted in more intelligibility breakdowns were content words. The words *slabs* and *scoop*, for instance, caused many intelligibility breakdowns, and many different lexical substitutions happened.

Even though both ASR programs had trouble transcribing the same words, some differences were also observed. The words *Bob* and *scoop* caused more problems for VoiceNotebook than for MW dictation while MW dictation was less accurate than VoiceNotebook only in transcribing the word *peas*. Analyzing the control group (n=5), VoiceNotebook also applied an incorrect strategy transcribing the first four items of the table among the native speakers' samples. The words were transcribed as *Labs* and *fix* (*eng200*), and *pins* and *Bop* (*eng10*) respectively. For MW dictation, on the other hand, only the word *and* was mistranscribed as *in* (*eng200*). This shows that, even for native speakers, specific lexical items might cause intelligibility breakdowns. Nevertheless, the two programs had trouble transcribing only two native speakers (*eng200* and *eng10*). The speaker *eng200* particularly showed a high speech rate. It demonstrates that a fluent speech might represent an extra barrier for the ASR programs.

Table 4. Items that caused ten or more intelligibility breakdowns

VoiceNotebook (*n*=35)			MW dictation (*n*=35)		
Item	TIB	Frequent Lexical Substitutions*	Item	TIB	Frequent Lexical Substitutions*
slabs	24	flags (4), laps (3)	*thick*	14	things (3), six (3)
thick	23	six (7), fix (5)	*spoons*	12	pounds (8)
spoons	22	pounds (4)	*and*	11	in (9)
Bob	13	but (9)	*peas*	10	peace (4), piece (3)
and	12	in (10)	*slabs*	10	-
scoop	11	-			

Source: authors

Data from all the samples, including native speakers. TIB = total number of intelligibility breakdowns. *Substitution strategy employed three or more times for that item.

Some patterns can be observed in the strategies employed by the programs. Many transcription mistakes occurred in both ASR programs for the same speech sample, meaning that speaker's output was too deviant for either of the dictation programs to identify as the intended text. Bearing this in mind, Table 5 below presents the most common lexical items that were wrongly transcribed for the two dictation programs separately. Table 5 also separates the EIL speakers into two L1 groups and lists the most common substitutions for each item. Looking at the possible causes for the substitutions, at least three could be identified: segmental, suprasegmental, and syntax-semantic, even though in some cases it is possible to identify more than one cause for the substitution.

Table 5. Items that caused five or more intelligibility breakdowns for the two EIL groups separated by the programs

	VoiceNotebook			MW dictation		
	Item	TIB	FLS	Item	TIB	FLS
L1 BP speakers (*n*=15)	*spoons*	12	-	*and*	7	in (7)
	thick	11	six (7)	*spoons*	6	pounds (4)
	slabs	11	flags (4)			
	and [2]	9	in (9)			
	scoop	5	-			
	the	5	a (5)			

	VoiceNotebook			MW dictation		
	Item	TIB	FLS	Item	TIB	FLS
L1 Sp speakers (*n*=15)	*slabs*	12	-	*thick*	10	things (3)
	thick	11	fix (3)	*slabs*	9	-
	Bob	9	but (7)	*peas*	8	piece (3)
	spoons	8	-	*spoons*	6	pounds (4)
	red bags	7	___ box (6)	*frog*	6	-
	bags	6	back(s) (5)	*five*	5	-
	scoop	6	-			
	peas	6	-			
	red	5	right (3)			
	a snack	5	it's not (4)			

Source: authors

TIB = total number of intelligibility breakdowns. FLS = Frequent Lexical Substitutions (substitution strategy employed three or more times for that item).

Substitution at the segmental level appeared to be partially due to the EIL speakers' L1 inventories. The word *thick* (/θɪk/) was frequently erroneously transcribed by both programs for both EIL speaker groups. The voiceless interdental fricative is not present as a phoneme in the consonantal inventories of the two languages in question. Nevertheless, the two EIL groups appeared to employ different consonantal substitutions as for the L1 Sp speakers the most frequent transcription was *fix* and for the L1 BP speakers, the most frequent transcription was *six*. Final voicing caused problems for the L1 Sp speakers resulting in intelligibility breakdowns. This became evident in the transcription of the words *peas* (/piz/) as *piece* or *peace* and *bags* (/bægz/) as *back(s)*, indicating inaccurate production of the final voiced consonant. The L1 BP speakers did not present this problem, probably because [z] is an allophone of [s] in Brazilian Portuguese. Similarly, the pronunciation of the word *five* by the L1 Sp speakers resulted in intelligibility breakdowns for MW dictation, possibly because it contains the phoneme /v/, not present in this L1s' inventory.

Intelligibility breakdowns due to vocalic deviations were obvious in the transcription (*these* as *this*, *store* as *star*, or *scoop* as *skip*). The vowel quality (e.g., /æ/ in *slabs*) and duration (e.g., /u/ in *scoop*) of some English vowels represented an extra difficulty for the speakers. Additionally, the word *Bob* was frequently transcribed as *but* for the L1 Sp speakers, indicating a possible problem with the vowel, even though a syntactic cause could also be possible: the program might not have expected a proper noun in that syntactic position, or the speech's prosody suggested a continuation of the speech, making the word *but* more probable.

Problems with phonotactics were also a common cause of intelligibility breakdowns evidenced in the transcription of consonant clusters as two separate words (*snow* as *is no*, *slabs* as *is left*, *a snack* as *is not*) or as different words (*Stella* as *Estella*) indicating the addition of an epenthetic vowel. Overall, s-cluster words resulted in many intelligibility breakdowns, as evidenced by their high occurrence in Table 5, indicating transfer of L1 phonotactic rules. Difficulties in stressing words also appeared to contribute to intelligibility issues as evidenced by [ˈrɛd ˌbɑks] being interpreted as *Redbox* rather than *red bags*, or [ˈʃikən] interpreted as *chicken* rather than *she can*.

Turning our focus to the suprasegmental features of the speech and connected speech phenomena, it could be seen that function words and their weak forms caused many intelligibility breakdowns even in fluent speakers. For example, the word *and* transcribed as *in* or *a* transcribed as *I* suggests that the programs had problems in identifying weak forms. Most problems related to connected speech occurred in both programs, but with different speakers regarding different lexical items. For example, connecting a word with the pronoun *her* was troublesome for both

programs (e.g., *ask her*). Moreover, the word *and* in the sentence *a small plastic snake and a big toy frog for the kids* caused many intelligibility breakdowns for both programs, especially for the L1 BP speakers.

Another possible cause for the substitutions employed by the ASR programs is related to syntax-semantic issues. For example, the word *scoop* caused intelligibility issues in the speech of *sp168* leading to the misinterpretation of *she can scoop* to *she kind of school*, where *of* was added by the program to accommodate for the misinterpretation of the verb into a noun. Similarly, *six spoons of fresh snow peas* was very hard to be interpreted, mainly for VoiceNotebook. This item caused varied lexical substitutions. A human listener would not understand the utterance as *six pounds of fresh nose piece* as MW dictation mistranscribed the speech of *sp109*. Nonetheless, *six pounds* was a very common alternative to this item on MW dictation, which respects both syntactic and semantic contexts. In general, MW dictation seems to respect semantics more than VoiceNotebook does once the lexical substitutions are more coherent to the context. For example, VoiceNotebook transcribed the passage *a big toy frog for the kids* as *a big tree frog for the kids* on some occasions. Furthermore, function words occasionally caused intelligibility breakdowns, especially the definite article *the*, being transcribed into the indefinite article. Once they occupy the same syntactic position, both words are plausible.

All in all, both programs had more difficulties transcribing Sp speakers than BP speakers. Segmental and suprasegmental issues resulted in intelligibility breakdowns as did syntactic and semantic grouping. Furthermore, fluent speech may cause intelligibility breakdowns, mainly due to the connected speech phenomena. Moreover, intelligibility breakdowns caused by segmental features were more common in VoiceNotebook. It seems that MW dictation relies more on the context; thus disregarding some pronunciation deviations. Notwithstanding, the pronunciation deviations that caused intelligibility breakdowns for the ASR-based dictation programs varied according to the L1 and the program, happening in both segmental and suprasegmental levels.

6. DISCUSSION

The results indicated that overall, the EIL speakers were highly intelligible to the two programs. On average, VoiceNotebook correctly transcribed the EIL speech in 80% of the cases, and MW dictation correctly transcribed the EIL speech in 89% of the cases. Results revealed that this difference was statistically significant. In other words: MW dictation provided more accurate transcripts of the EIL speakers' speech in our study. The intelligibility score obtained by MW dictation is congruent with the results obtained from McCrocklin and Edalatishams (2020) using Google Voice Typing to transcribe L1 Chinese (90,99%) and L1 Spanish speakers (92,73%).

Even though the EIL speakers' intelligibility was lower than that of the native English speakers (VN M=95.36%, MW M=98.33%), native speech was not a guarantee for perfect intelligibility as testified by the range of native speaker intelligibility scores (85-100%). This is not necessarily surprising if we consider previous research on intelligibility. Even though it is known that the amount of segmental errors present in the L2 speech is associated with decreased intelligibility (e.g., NAGLE; HUENSCH, 2020), other factors, such as noise (MUNRO, 1998) affect intelligibility as well. If we observe more in detail the native English speaker who obtained the lowest intelligibility score (85% by VoiceNotebook), it can be seen that out of the ten mistakes that were present in the transcript, six corresponded to missing words. The speaker's speech rate was rather fast, and it thus seems that as with L2 listeners (KIVISTÖ DE SOUZA; MORA, 2012), speech rate can also affect intelligibility assessed by ASR.

The analyses also indicated that the L1 BP speakers (M=92.80%) were more intelligible to MW dictation than the L1 Sp speakers (M=85.19%). Even though a similar pattern was observed with VoiceNotebook, the difference between L1 BP (M=83.59%) and L1 Sp speaker intelligibility (M=76.76%) did not reach statistical significance. We took specific care in controlling for the linguistic variables when selecting participants for the two EIL groups, and statistically, the groups did not differ in terms of AOL, LOR, or L2 experience. Nonetheless, it should be borne in mind that only a very limited amount of data in relation to individual variables were available from the speakers and they were based on self-reports. Even though using an existing database provided access to speakers we otherwise would not have had access to due to the Covid-19 pandemic, it also brought some limitations to the study. Consequently, the current study does not provide enough evidence to determine whether VoiceNotebook and MW dictation judge speakers from different L1 backgrounds to be more intelligible than others. From studies with human judges, it is known that accent familiarity plays a role in intelligibility: listeners who are more familiar

with foreign-accented speech, tend to judge the L2 speech as more intelligible (e.g., DERWING; MUNRO, 1997; NAGLE; HUENSCH, 2020). Even though it is beyond the scope of this paper and acknowledging that ASR usually uses native speaker speech samples as models (ROGERSON-REVELL, 2021), it would be interesting to know whether the developers of these ASR programs used L2 accented speakers to feed their database.

The results showed that not only segmental deviations but also suprasegmental deviations, such as syllable structure and vowel reduction led to decreased intelligibility. Difficulties in stressing words also appeared to contribute to intelligibility issues, which is not surprising considering that human listeners also rely heavily on suprasegmental features in identifying words from a speech stream (JOHNSON; JUSCZYK, 2001) and that non-target-like word stress in L2 English can be detrimental to intelligibility (ZIELINSKI, 2008).

Analyzing how ASR intelligibility compares to actual communication situations between language users is not the scope of this paper, but it is tempting to hypothesize how they would differ and in which cases the ASR program might present some shortcomings. In communication between language users, context usually aids the listener when a pronunciation deviation is about to lead to communication breakdowns. Previous studies show that semantic context, familiarity with the topic, and the speaker are positively related to intelligibility (e.g., GASS; VARONIS, 1984; KENNEDY; TROFIMOVICH, 2008)[7]. We assume that some intelligibility issues faced by the ASR programs in the present study probably would not have happened in real life communication. For example [ðɪz θɪŋz] was in some cases transcribed as *this things* whereas in communication between language users, the context would be likely to provide the intended message *these things* despite the inaccurate vowel production. A similar phenomenon was observed in cases such as *Fresno Ps* for *fresh snow peas*. Nevertheless, we can observe that the programs under analysis did attempt to use syntactic and semantic context to make sense of the message when intelligibility was endangered. This becomes obvious in cases where the programs, especially VoiceNotebook, added words to the transcript that were not present in the acoustic signal (e.g., *she kind of school* instead of *she can scoop*).

Overall, the data indicated that for these EIL speakers, MW dictation was more accurate in transcribing their speech as intended. Consequently, transcripts performed by VoiceNotebook evidenced a higher use of compensatory strategies. When the message was not understood, both programs' most frequent strategy was to substitute the intended word with another lexical item (*substitution*). The order of the other strategies varied slightly between the two programs. The second most used strategy for MW dictation was substituting two words with one lexical item (e.g., *we also* transcribed as *Wilson*), whereas VoiceNotebook employed omission. The higher use of omission by VoiceNotebook than MW dictation was statistically significant. MW dictation rarely missed more than one word in a row in the transcription process whereas for VoiceNotebook omission of entire clauses took place occasionally. Furthermore, the addition of words absent from the speech signal was also significantly more frequent in VoiceNotebook than MW dictation. This might also happen because VoiceNotebook seemed to be more susceptible to pronunciation deviations on a segmental level. It was possible to notice that, mainly for VoiceNotebook, those deviations interfere more than the semantic context.

From the point of view of assessing the efficiency of the ASR-based dictation tools, the omission of words is problematic. First, if the reader of the transcript is not familiar with the original passage, they might not know that some words were omitted as the programs do not indicate missing words in any way. Second, with omission, the reader also receives no clues as to the phonological content of the omitted part, contrary to when an item is substituted with another lexical item (e.g., *keys* instead of *kids*). However, what is problematic from a technical point of view, could be potentially beneficial for the learner. If the learner sees that the ASR program has missed many of the words that were uttered, it could lead to a reflection of the characteristics of the learner's output and offer opportunities for noticing the gap. Of course, the learner will not know whether the omissions have occurred because of unintelligible speech segments or a system limitation. Just like with studies focusing on feedback (e.g., LYSTER; SAITO, 2010), it could be interesting to study how language users react to the different compensatory strategies used by the ASR-based dictation tools and if one leads into more noticing of the gap than others.

From a pedagogical point of view, some issues emerged. Regarding the selection of which ASR-based dictation tools should be used, an interesting strategy might be to explore which of them offers a more suitable intelligibility score according to the learner's L1. Moreover, it is prudent to select an appropriate passage to be transcribed

7. Notice however opposing view from Jenkins (2002) who argues that lower to intermediate proficiency level language learners are not able to use contextual information into their advantage to the same extent as native speakers or fluent bilinguals.

containing common words and phrases, besides avoiding proper nouns and unexpected combinations of words (e.g., *five thick slabs of blue cheese, six spoons of fresh snow peas*) to circumvent the limitations of the programs. This recommendation is also supported by Ashwell and Elam (2017) who found that certain collocations and proper nouns caused problems for the ASR system. Hence, understanding which items are difficult for the system to recognize is advisable to tailor more appropriate ASR-based pronunciation practices. Lastly, it is important to bear in mind that the accuracy of ASR programs is affected by the output device and the environment. Therefore, a proper microphone, a quiet place, and a stable internet connection are of paramount importance to achieve better results using these tools.

Although the ASR programs did not present an intelligibility score equal to 100%, which is expected since "ASR will never be 100% accurate" (KNILL et al., 2018, p. 1641), they suggest interesting opportunities for pronunciation teaching. As Yoshida (2018) points out, ASR programs would not be of great pedagogical use if they transcribed nonnative speech without any intelligibility breakdowns, disregarding all pronunciation deviations. Thus, what initially might be considered as a limitation of the programs may represent an opportunity for the learners to be aware of their pronunciation difficulties and therefore mold their learning goals according to them. This extended practice is particularly relevant when practicing segmental features which are not present in the L1s' inventory.

ASR-based dictation tools do not offer specific feedback on how to produce a more intelligible speech. Therefore, dictation tools are ideal as an additional learning opportunity but not as the sole source of pronunciation learning as learners will also need explicit instruction and feedback to improve their pronunciation. Moreover, ASR-based dictation tools have the potential to be facilitative of autonomous learning; however, additional guidance from the instructor might be required. For example, the connected speech phenomena caused many intelligibility breakdowns. Thus, fluent and proficient speech may represent an extra barrier for the ASR programs due to the speech rate. Then, respecting the limitations of the programs is required. Moreover, MW dictation could deal with connected speech more appropriately than VoiceNotebook, possibly being a better choice for more proficient learners. On the other hand, VoiceNotebook focused more on the acoustic signal and relied more on the segmental features, being a convenient choice for minimal pair practice, for instance. Nevertheless, learners can combine the usage of both tools as a way of practicing the intended passage and comparing the accuracy of both programs; thus, practicing twice. To summarize, ASR-based dictation tools might be a more suitable option for accuracy practice instead of fluency.

CONCLUSIONS

This study set to examine the intelligibility of EIL speakers from two L1 backgrounds as assessed by two ASR-based dictation programs. Our objective was to determine how well the programs could understand EIL speakers and what strategies the programs employed when intelligibility was compromised. We also set to examine which pronunciation deviations were likely to cause intelligibility breakdowns and what lexical items in the elicitation paragraph were not understood most frequently. Results indicated that the two ASR programs understood EIL speakers' speech fairly well. We thus concluded that the two ASR-based dictation tools filled the basic requirements needed for language learners to employ them as additional out-of-the-classroom pronunciation learning tools.

Learners may benefit from these tools as they offer extended opportunities to practice both segmental and suprasegmental features of the target language. This is even more relevant in the Latin American context, where contact with native speakers of English is scarce. In addition, pronunciation practice tends to be neglected inside the language classroom due to time constraints (MUNRO; DERWING, 2015). Thus, ASR-based dictation tools might be a suitable resource for this end. Furthermore, ASR-based dictation practice is available whenever the learner has the chance to be online and in a quiet place; therefore, these tools offer endless opportunities for practice. Nonetheless, as the present study demonstrates, the proper guidance of a teacher is rather necessary to circumvent the limitations of the programs, tailor the passages according to the learner's needs and proficiency level, and provide appropriate feedback on the intelligibility breakdowns indicated by the programs. In short, even though ASR-based dictation tools are not a substitute for language teachers, they may be adequate for autonomous learning.

The present study presents some shortcomings that limit the amount of generalizations that can be drawn from it. As mentioned previously, we did not have direct access to the speakers who provided the speech samples and thus could not control for the recording conditions, the type of background information collected, or the speech

elicitation material. As for the speech elicitation paragraph, whereas it elicited a wide variety of English phones, its semantic content is somewhat strange, presenting low-frequency words (e.g., *scoop*) and combinations of phrases (e.g., *five thick slabs of blue cheese*) that are not necessarily representative of language learners' everyday language use. The speakers might not have been familiar with some of the low-frequency items of the paragraph, resulting in atypically less target-like production as lexical knowledge is tightly linked to L2 speech learning (MORA, 2005). Finally, it can be argued whether paragraph reading can be representative of real-life communicative situations and the demands posed by online language processing. Future studies should look at how ASR deals with spontaneous speech samples. Another fruitful line of research would be to examine how ASR-based dictation tools compare to human listeners in terms of intelligibility. Should ASR-based dictation tools assess intelligibility similarly to human listeners, many practical applications in the field of language testing and teaching could rise. For example, teachers (and examiners) could outsource intelligibility assessment for ASR and focus on other aspects of L2 speech learning that receive too little attention.

ASR-based dictation tools can be a helpful additional pronunciation learning tool for L2 users. For some learners, the feedback offered by them might be less face-threatening than feedback from other language users. ASR-based dictation tools can increase noticing of the gap, but it is unclear yet whether language users can actually use the feedback to improve their intelligibility. Keeping in mind the notoriously short time for pronunciation instruction in language curriculums, we nevertheless believe that any aid is welcomed.

CONFLICT OF INTEREST

The authors declare that they have no affiliations or involvement with institutions that may have financial or non-financial interests with the subject matter discussed in the article.

AUTHOR CONTRIBUTION STATEMENT

Hanna Kivistö de Souza created the task, selected the speakers, analyzed quantitative data and contributed to the writing of the manuscript.

William Gottardi performed the data collection, analyzed qualitative data and contributed to the writing of the manuscript.

DATA AVAILABILITY STATEMENT

The data used in the study were obtained from the Speech Accent Archive - available at https://accent.gmu.edu/. Information about the specific speakers that were chosen for the analyses can be obtained from the first author.

REFERENCES

ASHWELL, T.; ELAM, J. R. (2017). How accurately can the google web speech API recognize and transcribe Japanese L2 english learners' oral production? *JALT CALL Journal*, v. 13, n. 1, p. 59-76.

CARLET, A.; KIVISTÖ DE SOUZA, H. (2018). Improving L2 pronunciation inside and outside the classroom: Perception, production and autonomous learning of L2 vowels. *Ilha do Desterro*, v.71, n.3, p.99-123.

BASHORI, M. et al. (2020). Web-based language learning and speaking anxiety. *Computer Assisted Language Learning*, v. 0, n. 0, p. 1-32.

BASHORI, M. et al. (2021). Effects of ASR-based websites on EFL learners' vocabulary, speaking anxiety, and language enjoyment. *System*, v. 99, n. April, p. 102496.

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

CUCCHIARINI, C.; NERI, A.; STRIK, H. (2009). Oral proficiency training in Dutch L2: The contribution of ASR-based corrective feedback. *Speech Communication*, v. 51, n. 10, p. 853-863.

CUCCHIARINI, C.; STRIK, H. (2018). Automatic Speech Recognition for second language pronunciation training. In: *The Routledge handbook of contemporary English pronunciation*. Routledge. p. 556-569.

DERWING, T. (2010). Utopian goals for pronunciation teaching. (J. Levis, K. LeVelle, Eds.) In: 1st Pronunciation in Second Language Learning and Teaching Conference. Proceedings... Ames, IA: Iowa State University.

DERWING, T.; MUNRO, M. (1997). Accent, intelligibility, and comprehensibility: Evidence from Four L1s. *Studies in Second Language Acquisition*, v. 19, n. 1, p. 1-16.

DIZON, G.; TANG, D. (2020). Intelligent personal assistants for autonomous second language learning: An investigation of Alexa. *JALT CALL Journal*, v. 16, n. 2, p. 107-120.

GASS, S.; VARONIS, E. M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language learning*, v. 34, n. 1, p. 65-87.

GOLONKA, E. M. et al. (2014). Technologies for foreign language learning: A review of technology types and their effectiveness. *Computer Assisted Language Learning*, v. 27, n. 1, p. 70-105.

HENRICHSEN, L. E. (2020). An Illustrated Taxonomy of Online CAPT Resources. *RELC Journal*, 52(1), 179-188.

INCEOGLU, S.; LIM, H.; CHEN, W. H. (2020). Asr for EFL pronunciation practice: Segmental development and learners' beliefs. *Journal of Asia TEFL*, v. 17, n. 3, p. 824-840.

JENKINS, J. (2002). A sociolinguistically based, empirically researched pronunciation syllabus for English as an international language. *Applied linguistics*, v. 23, n. 1, p. 83-103.

JENKINS, J.; COGO, A.; DEWEY, M. (2011). Review of developments in research into English as a lingua franca. *Language teaching, 44*(3), 281-315.

JOHNSON, E.; JUSCZYK, P. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language*, v. 44, p. 548-567.

JURAFSKY, D.; MARTIN, J. H. (2021). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (3rd ed., Unpublished Manuscript). Available at: <https://web.stanford.edu/~jurafsky/slp3/ed3book_sep212021.pdf>. Accessed: Nov, 1st 2021.

KENNEDY, S.; TROFIMOVICH, P. (2008). Intelligibility, comprehensibility, and accentedness of L2 speech: The role of listener experience and semantic context. *Canadian Modern Language Review*, v. 64, n. 3, p. 459-489.

KIM, I. S. (2006). Automatic speech recognition: Reliability and pedagogical implications for teaching pronunciation. *Educational Technology and Society*, v. 9, n. 1, p. 322-334.

KIVISTÖ DE SOUZA, H.; MORA, J. C. Speech rate effects on L2 vowel production and perception. In: CELSUL, 2012, Cascavel, Paraná. Anais do X Encontro do CELSUL-Círculo de Estudos Linguísticos do Sul. Cascavel, 2012.

KNILL, K. M. et al. (2018). Impact of ASR performance on free speaking language assessment. In: Annual Conference of the International Speech Communication Association, INTERSPEECH. Proceedings… v. 2018- Septe, p. 1641-1645.

LEVIS, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *Tesol Quarterly*, v. 39, n. 3, p. 369-377.

LEVIS, J.; SUVOROV, R. (2013). Automatic Speech Recognition. In: CHAPELLE, C. A. (Ed.). *The encyclopedia of applied linguistics*. New York: Wiley-Blackwell. p. 316-323.

LIAKIN, D.; CARDOSO, W.; LIAKINA, N. (2015). Learning L2 pronunciation with a mobile speech recognizer: French/y/. *CALICO Journal*, v. 32, n. 1, p. 1-25.

LIAKIN, D.; CARDOSO, W.; LIAKINA, N. (2017). Mobilizing Instruction in a Second-Language Context: Learners' Perceptions of Two Speech Technologies. *Languages*, v. 2, n. 3, p. 11.

LYSTER, R.; SAITO, K. (2010). Oral feedback in classroom SLA: A Meta-Analysis. *Studies in Second Language Acquisition*, v. 32, n. 2, p. 265-302.

MCCROCKLIN, S. (2019a). Dictation programs for second language pronunciation learning: Perceptions of the transcript , strategy use and improvement. v. 7, n. 2, p. 137-157.

MCCROCKLIN, S.; EDALATISHAMS, I. (2020). Revisiting Popular Speech Recognition Software for ESL Speech. *TESOL Quarterly*, v. 54, n. 4, p. 1086-1097.

MCCROCKLIN, S. M. (2014). Dictation programs for pronunciation learner empowerment. In: 5th pronunciation in second language learning and teaching conference. Proceedings… n. September, p. 30-39.

MCCROCKLIN, S. M. (2016). Pronunciation learner autonomy: The potential of Automatic Speech Recognition. *System*, v. 57, n. April 2016, p. 25-42.

MICROSOFT. (2021). Dictate Your Documents in Word. Available at: <https://support.microsoft.com/en-us/office/dictate-your-documents-in-word-3876e05f-3fcc-418f-b8ab-db7ce0d11d3c#Tab=Web>. Accessed: Nov, 1st 2021.

MORA, J. C. (2005). Lexical knowledge effects on the discrimination of non-native phonemic contrasts in words and nonwords by Spanish/Catalan bilingual learners of English. In: ISCA Workshop on Plasticity in Speech Perception.

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

MUNRO, M. (1998). The effects of noise on the intelligibility of foreign-accented speech. *Studies in Second Language Acquisition*, v. 20, n. 2, p. 139-154.

MUNRO, M. J.; DERWING, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language learning*, v. 45, n. 1, p. 73-97.

MUNRO, M. J.; DERWING, T. M. (2015). Intelligibility in research and practice: Teaching priorities. In: REED, M.; LEVIS, J. M. (Eds.). *The Handbook of English Pronunciation*. Wiley Online Library. p. 375-396.

NAGLE, C. L.; HUENSCH, A. (2020). Expanding the scope of L2 intelligibility research: Intelligibility, comprehensibility, and accentedness in L2 Spanish. *Journal of Second Language Pronunciation*. 6.

NERI, A.; CUCCHIARINI, C.; STRIK, H. (2006). ASR-based corrective feedback on pronunciation: Does it really work? *INTERSPEECH 2006 and 9th International Conference on Spoken Language Processing, INTERSPEECH 2006 - ICSLP*, v. 4, n. May 2014, p. 1982-1985.

NERI, A.; CUCCHIARINI, C.; STRIK, H. (2008). The effectiveness of computer-based speech corrective feedback for improving segmental quality in L2 Dutch. *ReCALL*, v. 20, n. 2, p. 225-243.

ROGERSON-REVELL, P. M. (2021). Computer-Assisted Pronunciation Training (CAPT): Current Issues and Future Directions. *RELC Journal*, v. 52, n. 1, p. 189-205.

VOICENOTEBOOK. (2021). Voice Notebook Homepage. Available at: <https://voicenotebook.com>. Accessed: Nov, 1st 2021.

WEINBERGER, S. (2015). *Speech Accent Archive*. George Mason University. Available at: <http://accent.gmu.edu>. Accessed: Nov, 1st 2021.

YOSHIDA, M. T. (2018). Choosing technology tools to meet pronunciation teaching and learning goals. *The CATESOL Journal*, v. 30, n. 1, p. 195-212.

YU, D.; DENG, L. (2015). *Automatic Speech Recognition A Deep Learning Approach*. London: Springer.

ZIELINSKI, B. W. (2008). The listener: No longer the silent partner in reduced intelligibility. *System*, v. 36, n. 1, p. 69-84.

Recebido: 28/3/2022
Aceito: 27/9/2022
Publicado: 30/9/2022

Este preprint foi submetido sob as seguintes condições:

- Os autores declaram que estão cientes que são os únicos responsáveis pelo conteúdo do preprint e que o depósito no SciELO Preprints não significa nenhum compromisso de parte do SciELO, exceto sua preservação e disseminação.
- Os autores declaram que os necessários Termos de Consentimento Livre e Esclarecido de participantes ou pacientes na pesquisa foram obtidos e estão descritos no manuscrito, quando aplicável.
- Os autores declaram que a elaboração do manuscrito seguiu as normas éticas de comunicação científica.
- Os autores declaram que os dados, aplicativos e outros conteúdos subjacentes ao manuscrito estão referenciados.
- O manuscrito depositado está no formato PDF.
- Os autores declaram que a pesquisa que deu origem ao manuscrito seguiu as boas práticas éticas e que as necessárias aprovações de comitês de ética de pesquisa, quando aplicável, estão descritas no manuscrito.
- Os autores declaram que uma vez que um manuscrito é postado no servidor SciELO Preprints, o mesmo só poderá ser retirado mediante pedido à Secretaria Editorial do SciELO Preprints, que afixará um aviso de retratação no seu lugar.
- Os autores concordam que o manuscrito aprovado será disponibilizado sob licença Creative Commons CC-BY.
- O autor submissor declara que as contribuições de todos os autores e declaração de conflito de interesses estão incluídas de maneira explícita e em seções específicas do manuscrito.
- Os autores declaram que o manuscrito não foi depositado e/ou disponibilizado previamente em outro servidor de preprints ou publicado em um periódico.
- Caso o manuscrito esteja em processo de avaliação ou sendo preparado para publicação mas ainda não publicado por um periódico, os autores declaram que receberam autorização do periódico para realizar este depósito.
- O autor submissor declara que todos os autores do manuscrito concordam com a submissão ao SciELO Preprints.