Language Studies Dossier: Article

DOI: http://dx.doi.org/10.5007/2175-8026.2024.e98870

EXAMINING VOWEL INTELLIGIBILITY IN BRAZILIAN PORTUGUESE EFL LEARNERS: A STUDY USING AUTOMATIC SPEECH RECOGNITION

Fhelippe Waltelon Souza dos Santos[1*]

Hanna Kivistö de Souza[1**]
[1]Universidade Federal de Santa Catarina, Florianópolis, SC, Brasil

Abstract

As traditional EFL classrooms often struggle with limited time and input for pronunciation instruction, technologies such as Automatic Speech Recognition (ASR) may serve as a complementary tool to aid learners in the acquisition of specific English sounds. However, in order to consider the usefulness of ASR tools, more careful testing should be conducted to observe their ability to perform in specific contexts. The objective of the present study was to replicate Kivistö-de Souza and Gottardi's experiment (2022) by examining how *Microsoft Word Dictate* judged the intelligibility of Brazilian Portuguese speakers of English. Contrary to Kivistö-de Souza and Gottardi (2022), who examined continuous speech, the present study focused on vowels embedded in isolated words. Results showed that the intelligibility was low (50%) in comparison to the scores presented in Kivistö-de Souza and Gottardi (2022). Discussion centers around the uses of ASR on EFL pronunciation development in autonomous learning and instructed settings.

Keywords: English pronunciation; Automatic speech recognition; Intelligibility; FL vowel production; Autonomous learning.

[*] MA in English Linguistic and Literary Studies (UFSC), PhD candidate in Programa de Pós-graduação em Inglês, UFSC. Email: fwsdsantos@gmail.com. ORCID https://orcid.org/0000-0002-8315-2098.

[**] PhD in Applied Linguistics (University of Barcelona). Professor at Departamento de Língua e Literaturas Estrangeiras (DLLE/PPGI) at Universidade Federal de Santa Catarina. E-mail: hanna.souza@ufsc.br. Orcid: https://orcid.org/0000-0002-8498-2691 .

1. Introduction

Acquiring English as a Foreign Language (EFL) pronunciation in instructed settings appears to be an arduous task for many non-native speakers (Celce-Murcia et al., 2010). Among other factors, this stems from time restrictions, inadequacies in the quantity and quality of the input, the overall difficulty of acquiring specific phonological features, and the learners' emotional states regarding their pronunciation development (Baran-Łucarz, 2016; Silveira, 2004). In particular, affective factors appear to significantly impact non-native speech production. Anxiety over the fear of pronouncing the target language incorrectly around teachers and peers can effectively exclude certain speakers from partaking in speaking opportunities in the classroom (Phillips 1992; Price 1991; Young, 1992 as cited in Baran-Łucarz, 2014).

It is because of these existing obstacles in EFL pronunciation learning in instructed settings, that some researchers have turned into investigating speech technologies as a way to provide learners with more exposure to English sounds through perception and pronunciation practice (Chun, 2011, 2013; Tejedor-Garcia et al., 2020; Thomson & Derwing, 2014). In particular, ASR appears to be one of the most commonly applied technologies in EFL pronunciation studies, specifically due to its ability to provide learners with an autonomous and personalized experience, that permits EFL learners to become more aware of their own intelligibility issues, especially when the ASR output does not match the learner's intended message (Mroz, 2018). Additionally, ASR pronunciation practice can be carried out in an environment chosen by the learners, enabling them to practice without social risks, such as the fear of negative evaluation.

Even though ASR technologies have improved in their understanding of non-native speakers, further analyses need to be conducted with non-native speech in specific contexts in order to assess whether ASR tools can be autonomously used by EFL learners for pronunciation development. As an attempt to contribute to the field of EFL pronunciation acquisition and speech technologies, this study aims to replicate Kivistö-de Souza and Gottardi's (2022) study by addressing some of its limitations. Specifically, this study applies the same approach to EFL intelligibility assessment by presenting a specific ASR tool (*Microsoft Word Dictate)* with English speech by L1 Brazilian Portuguese (BP) speakers. However, contrary to the Kivistö-de Souza and Gottardi's (2022) study, which examined intelligibility at a sentence level and employed a generic pre-existing paragraph reading task, the present paper examines intelligibility of specific English vowels (/i ɛ æ ɑ u ʊ/). Our objective was to assess whether presenting EFL speech in isolated form (vs. continuous speech) would result in more accurate assessments (measured as vowel intelligibility) by the ASR tool. We also examined whether the compensatory strategies employed by the ASR tool in cases of intelligibility breakdowns would differ from those employed in continuous speech. We expect the findings to contribute to the discussion on how ASR tools can be efficiently employed in EFL pronunciation development.

2. Background to the study

In this section we discuss previous research on the areas pertaining to the study: EFL speech intelligibility, the acquisition of English vowels by BP speakers and automatic speech recognition tools and their benefits for English learning.

2.1 Acquisition of English vowels by L1 BP speakers

In order to determine which difficulties may arise when acquiring English sounds, it is important to understand how to approach and measure English pronunciation. While the concept of native-like accuracy still exists and is used in specific contexts, as can be seen in mobile pronunciation instruction applications (e.g. Elsa), theory now steers towards a less idealized model of assessing EFL speakers. Since acquiring "native-like" pronunciation in a foreign language appears to be a difficult, often unattainable, goal for most EFL learners (Chun, 2013; Derwing, 2010), the focus has changed to evaluating the degree of pronunciation proficiency that is necessary for the learner to be intelligible in most situations. Munro and Derwing's (1995) approach to *Intelligibility* has been increasingly applied in studies, textbooks and classrooms (Derwing & Munro; 2022; Levis & Silpachai 2022). *Intelligibility*, according to Munro and Derwing (1999, p. 289), is "(...) the extent to which a speaker's message is actually understood by a listener", and as such it is frequently measured through the orthographic transcription of what the listener has heard (Gonçalves & Silveira, 2015; Kang et al., 2018; Miller, 2013; Munro & Derwing, 1995).

Some of the greatest challenges in English pronunciation acquisition are related to determining which sounds could be problematic for EFL speakers' intelligibility (Jin & Liu, 2014; Lima Junior, 2015, 2017; Munro & Derwing, 2006). Specifically, with *Intelligibility* being the rising approach regarding EFL pronunciation goals (Levis, 2005), it is important to consider which specific English sounds can cause significant intelligibility breakdowns. One of the factors contributing to intelligibility issues is the learners' difficulty in discerning between English sounds that are perceived similarly to L1 sounds (Flege & Bohn, 2021). Brazilian Portuguese (BP) speakers of English struggle especially with the accuracy and intelligibility of some English vowels (Silveira et al., 2009; Bion et al., 2006; Gonçalves & Silveira, 2015; Lima Junior, 2015, 2017, 2019; Nobre-Oliveira, 2007; Rauber, 2006).

These difficulties can be partly traced to the L1 vowel inventory: whereas in the BP vowel space there is one oral high front vowel phoneme (/i/), one oral high back vowel phoneme (/u/) and no low front vowels, in General American English, there are two high front vowels (/i-ɪ/), two high back vowels (/u-ʊ/) and one low front vowel (/æ/). Consequently, the BP speakers may create composite L1-FL phonetic categories by assimilating the two English phonemes into one category (Flege & Bohn, 2021). In this way, the BP learners of English may present difficulties in perceiving and producing the English /i ɪ u ʊ æ/ vowels, so that the

items in the minimal pairs *bat-bet*, *beat-bit* and *pool-pull* will be produced and heard as essentially the same lexical item. Lima Junior (2015) has researched the production and perception of English vowels by Brazilian Portuguese speakers extensively, evidencing that BP speakers, in fact, tend to assimilate the vowel pairs /i -ɪ/, /ɛ - æ/ and /u - ʊ/ into one vowel category. Consequently, Lima Junior claims that the vowels produced by the BP speakers "tend to occupy an intermediate position between the target vowels produced by the native speakers" (p. 26, our translation) when comparing native and non-native speakers of English.

However, all intelligibility and accuracy problems cannot be traced back to L1 vowel inventory. Gonçalves and Silveira (2015) looked into the difficulties faced by BP speakers of English regarding their production of English low-front vowels (/i - e/) by examining the roles of word frequency and familiarity. The authors report a positive correlation between intelligibility, word familiarity and word frequency, suggesting that intelligibility breakdowns may arise not only from processes of category assimilation, but also due to the presence of unfamiliar words.

Considering the difficulties L1 BP speakers have with English vowels, the question arises about how intelligible these speakers would be to current speech technologies, since ASR offers interesting possibilities for EFL pronunciation development (Thomson & Derwing, 2014; Mroz, 2018).

2.2 Automatic Speech Recognition and EFL speech learning

Automatic Speech Recognition is technology developed to transcribe verbal input into text, which is then digitally displayed via computer monitors and mobile devices (Levis & Suvorov, 2012). It can also be utilized for accessibility functions through intelligent personal assistants (IPAs), such as Amazon Echo and Google Assistant, which can aid users in navigating their mobile devices (setting alarms, navigating the internet), and home appliances (turning on and off lights and devices) through the use of voice commands.

While ASR technology has existed since the 1960s, its first implementation for FL learning was conducted in 1982 by Wohlert and his "German-By-Satellite" program (as cited by Wachowicz & Scott, 1999). In it, FL learners of German studied the language via satellite transmissions of lectures and practiced their pronunciation via their Apple II Plus device paired with a primitive ASR technology that could transcribe a small set of single words. Studies such as the one conducted by Wohlert evidence that the capabilities of ASR systems as a whole were relatively primitive until the early 2000s, when the technology was equipped with the ability to transcribe not only significantly larger amounts of vocabulary, but also continuous speech (Juang & Rabiner, 2005).

This is not to say, however, that ASR systems have been efficient and accessible for those using the technology in their second language. Coniam (1999) and Derwing (2000) claim that the technology of the time, specifically the ASR tool *Dragon NaturallySpeaking*, was significantly less accurate (approximately 20%) in transcribing non-native English speech, in contrast to its degree of accuracy

when transcribing speech uttered by native English speakers. That was because Dragon *NaturallySpeaking* was created as a speaker-dependent system, relying on individual training of the tool by a single speaker, instead of providing users with a pre-existing speech database in which the program is trained on, like the most modern ASR tools used today.

The modern speaker-independent systems function by comparing the user's spoken input to large databases of natural speech the tool has been trained with. Examples of modern tools that are categorized as such systems include *Google Voice* and Microsoft's *Azure AI* and, in particular, its speech recognition system. Nevertheless, these systems may encounter problems due to the type of data they are trained on, as EFL speech corpuses are still scarcely available (Shibano et al., 2021). This leads to the application of conversion techniques that adapt and normalize EFL speech samples into more intelligible samples according to the system itself.

Despite the limited number of EFL speech corpora, recent data on some speaker-independent systems such as *Google Voice* show promising results. As McCrocklin and Edalatishams (2020) state, this tool has been able to significantly close the gap between degree of accuracy of transcriptions between native and non-native speakers of English. McCrocklin and Edalatishams (2020) assessed how *Google Voice Typing* judges the productions of 30 participants (10 L1 English, 10 L1 Chinese Mandarin, 10 L1 Spanish) as opposed to human listeners. Each participant produced 60 sentences, and intelligibility judgements (by the 37 L1 English human listeners and by *Google Voice Typing*) were calculated via orthographic transcriptions. Overall, comparing the judgments by human listeners and the ASR software, it appeared that the gap in intelligibility was remarkably smaller than in previous studies (Coniam, 1999; Derwing, 2000), indicating that the ASR technology was falling behind only 5% in comparison to the human listener judgments for both Chinese and Spanish non-native speakers of English. Thus, non-native speakers of English in this study were only 5% less intelligible for *Google Voice Typing* than they were to human listeners.

Foreign language users can benefit from using ASR for EFL pronunciation development. The positive effects range from increased motivation and learner autonomy to lowered speaking anxiety (Bashori et al., 2021; Mroz, 2018). There are also ASR tools designed exclusively for EFL instruction that have shown positive results in increasing English pronunciation intelligibility (Burleson 2007; Hincks, 2003; Neri et al., 2008; Tejedor-Gárcia et al., 2020). Similarly, studies examining ASR software created for general purposes, (i.e., not for language learning), also suggest that positive gains for intelligibility, learner awareness and autonomy are possible (e.g., Mroz, 2018). For example, Tejedor-Gárcia and colleagues developed a computer assisted pronunciation tool (CAPT) that provided learners with five stages of pronunciation learning, including theory, exposure, discrimination, and pronunciation activities. This tool, which included ASR (pronunciation mode), was used subsequently in the same study to compare traditional in-class pronunciation learning to lab training with ASR and TTS technologies amongst 20 non-native English speakers. Results showed that the

CAPT condition reported for significant improvement in learners' production of English segments, akin to the in-class teaching condition.

Kivistö-de Souza and Gottardi (2022) also examined the modern ASR tools' ability to transcribe accented EFL speech by presenting paragraphs read by BP and Spanish speakers of English to two ASR dictation tools: Microsoft Word Dictate and VoiceNotebook. Their results indicated that the overall intelligibility of the non-native speakers was fairly high (80% for VoiceNotebook and 89% for Microsoft Word Dictation). When intelligibility breakdowns were encountered, i.e., when the ASR tools did not understand the speakers, the two ASR software employed different strategies (e.g. substituting the target word with another one, omitting an unintelligible word, adding more words to contextualize) in order to make sense of the message. The authors tracked some of the intelligibility breakdowns to inaccurate segmental production; vowel quality and duration (e.g. *these* as *this)*, consonant clusters (*snow* as *is no*) and connected speech phenomena (*she can* as *chicken*) contributed to the ASR tools' misunderstanding of the message.

Whereas Kivistö-de Souza and Gottardi's (2022) study shed some light on how ASR software behaves when presented with non-native speech, the study suffered from some limitations that would allow for more solid recommendations for the use of ASR as an EFL pronunciation development tool. As the speech samples came from an existing database (the Speech Accent Archive, Weinberger, 2015), the participants' language background or the elicitation material could not be controlled for. The paragraph the informants read included many low-frequency lexical items that might hinder the functioning of the ASR tool. The continuous nature of the paragraph reading also meant that the software presented some errors due to connected speech phenomena as well as the omission of some of the uttered words.

Thus, the objective of the present paper is, by replicating and addressing some of the shortcomings of Kivistö-de Souza and Gottardi (2022), to examine the suitability of Microsoft Word Dictate for the pronunciation practice of Brazilian speakers of English. We expect that by providing more data on how the ASR technology behaves when presented with Brazilian-accented English, more accurate recommendations can be made about the use of ASR for the English pronunciation development of this population.

Three research questions and hypotheses were posed:

RQ1: How does Microsoft Word Dictate judge the intelligibility of isolated words produced by BP learners of English?

> H1: Overall, intelligibility of isolated words is higher than intelligibility observed in Kivistö-de Souza and Gottardi (2022) as the transcriptions will not be affected by continuous speech phenomena, speed of delivery or potentially confusing contextual cues.

RQ2: How is the intelligibility of the target vowels /i ɛ ae ɑ u ʊ/ judged by Microsoft Word Dictate?

> H2: As Brazilian Portuguese EFL learners are likely to present inaccurate productions for these vowels (Lima Junior, 2015), intelligibility will be affected. As single L1 category assimilation is expected for /ɛ - æ/ and

/u - ʊ/ (Flege & Bohn, 2021), we hypothesize that this will be present in the learners' production and that especially /æ/ and /ʊ/ will have lower intelligibility as judged by Microsoft Word Dictate.

RQ3: What compensatory strategies does Microsoft Word Dictate use in cases of intelligibility breakdowns with isolated words?

H3: Following Kivistö-de Souza and Gottardi (2022), substituting the target word by another- most likely a minimal pair member- is expected to be the most frequently employed strategy when the learner's output is not transcribed accurately.

3. Method

The study followed a cross-sectional design in which BP speakers' vowel intelligibility was tested in one session through a production task, where isolated words containing the target vowels were read to the selected ASR tool (Microsoft Word Dictate). Participants additionally answered a linguistic background questionnaire.[1] Participants' intelligibility was assessed by examining the ASR transcription outputs for the presence of intelligibility breakdowns, which were quantified at the vowel and word level. Additionally, compensatory strategies used by the software in case of intelligibility breakdowns were inspected and compared with the results of Kivistö-de Souza and Gottardi (2022).

3.1 Participants

The participants were 18 English language and literature undergraduate students in a Brazilian Federal University enrolled at the time of the data collection in intermediate to advanced level English classes. Participants from these specific classes were invited to participate, as it was expected that the target-vowels would still pose a significant problem to EFL learners of this level. All participants were monolingual speakers of BP with self-assessed intermediate to advanced proficiency level in English. Information about participants' language background can be seen in Table 1.

Table 1 - *Participants' background*

Age	AOL	EFL experience	Time in English-speaking countries	Self-assessed English proficiency
24	13.11	12.9	0.33	3.71
(8.01)	(7.23)	(6.69)	(1.09)	(0.87)
18-51	0-32	7-34	0-1.5	2-5

Note. *N*=18. Age, AOL, EFL experience and Time in English-speaking countries expressed in years. Standard deviation between brackets. Range of responses on the final line. Self-assessed English proficiency was measured by calculating the mean of responses

to five proficiency categories (writing, reading, listening, speaking, and pronunciation), on a scale of 1-5 where 1 indicates beginner-level proficiency and 5 advanced-level proficiency. Age of Onset of Learning (AOL) refers to the first contact with English in an instructed setting. EFL experience = chronological age- AOL.

3.2 *Materials and procedures*

3.2.1 Stimuli

The vowels /i ɛ æ ɑ u ʊ/ were selected as target sounds taking into account previous research (e.g., Alves et al., 2014; Bion et al., 2006; Gonçalves & Silveira, 2015; Lima Junior, 2015, 2017, 2019; Nobre-Oliveira, 2007; Rauber, 2006) showing their difficulty for Brazilian Portuguese speakers. 30 carrier words were chosen containing the six target vowels. The selection criteria for the words were: monosyllabic and minimal pair or near-minimal pair words, in order to maintain a consistent phonetic contrast to minimize variability in the transcriptions given by the ASR tool. In specific, there were 30 CVC words (Table 2), out of which 18 ended in voiceless stops and 10 in voiced stops, with the two remaining words ending in the voiced alveolar lateral approximant /l/. In order to test the hypothesis in relation to the absence context (H1), the words were presented in isolation without a carrier sentence.

Prior to testing, the target words underwent small-scale piloting with Microsoft Word Dictate. The words were uttered with accurate and inaccurate vowel sounds by one of the researchers with the objective of verifying whether accurate and inaccurate transcriptions would result. For example, the target word *Peck* was produced inaccurately by either altering vowel quality or length, resulting in the transcription of the word *Pack*. Furthermore, for a target word to be included in this paper, they had to be possible to produce and correctly transcribed by *Microsoft Word Dictate* at least once by the researchers, which was the main purpose of the piloting procedures. On average, Microsoft Word Dictate reported behaved as expected by providing accurate transcriptions for accurate vowel productions and it was thus deemed appropriate for the purposes of the present study.

Table 2 - *Carrier words and target sounds*

Vowel	Carrier words				
i	Beat	Peak	Keep	Feed	Need
ɛ	Pet	Peck	Pep	Red	Head
æ	Bat	Rack	Rap	Sad	Bad
ɑ	Lot	Dock	Top	Rod	God
u	Boot	Luke	Loot	Pool	Dude
ʊ	Foot	Look	Took	Full	Hood

In the vowel production task, participants were presented with a PowerPoint presentation where each target word was shown orthographically one at a time

in a fixed order. Participants produced the item at their own pace and after a five second delay, the following item was presented.

3.2.2 Transcription Instrument

Kivistö-de Souza and Gottardi (2022) tested two ASR tools, out of which we chose the one that performed better in their study, i.e., provided more accurate transcriptions and less omissions. Thus, the ASR dictation feature found in Microsoft Word was chosen for this study. It is available for free through Word Online and utilizes Microsoft's Azure AI technology. Most importantly, it can be used through the Microsoft Word mobile application, which is how the participants were tested.

3.2.3 Procedures

Data were collected by testing participants individually at the university premises, in a silent room equipped with a projector and cellphone with an internet connection, as Microsoft Word accesses the cloud for the ASR feature. Participants saw each of the target words one by one through the projector and read them aloud. The order of the words was the same for all participants and each word remained on the screen for 5 seconds, after which the next word was presented. A cell phone was placed in front of the participants with the screen turned away from them and the participants were instructed to speak directly onto its built-in microphone. One of the researchers started a new paragraph after each utterance by manually pressing the button in the mobile device. This was done to prevent the ASR tool from picking up on contextual clues, as Microsoft Word Dictate attempts to guess words based on context and sentence structure. Participants did not see how their output was transcribed by the ASR tool. The program was configured to use American English voice recognition.

3.4 Data analysis

The same mixed-method approach as in Kivistö-de Souza and Gottardi (2022) was adopted with slight modifications. The transcription data were inspected for intelligibility breakdowns and coded. Intelligibility breakdowns were understood as cases in which Microsoft Word Dictate did not transcribe the word intended by the informant. The intelligibility breakdowns were computed by participant and by vowel in order to calculate participant intelligibility scores ((correctly transcribed words/total number of words)*100) and intelligibility scores for each target vowel (correctly transcribed words with the target vowel/total number of words with the target vowel)*100). Due to the nature of our data, we employed only four of the original five compensatory strategy categories of Kivistö-de Souza and Gottardi (2022). Compensatory strategies were understood as the ways in which the Microsoft Word Dictate behaved with intelligibility breakdowns. The four categories in the present study were substitution (substitution of the lexical item with a new one, e.g. *bat* instead of *bet*), omission (omission of the word present

in the recording), addition (inclusion of a word not present in the recording, e.g. *but bat* instead of *bat*) and 1x2 Substitution (substitution of one lexical item with two new ones, e.g. *do it* instead of *took*). As the informants produced isolated words, the 2x1 substitution category (substituting two words with one, e.g. *she can -> chicken*) from Kivistö-de Souza and Gottardi (2022) was not included in the analyses. Homophones (e.g. *Doc* instead of *Dock)* and singular forms instead of plural ones (e.g. *Boots* instead of *Boot*) were not considered errors.

Descriptive and inferential statistics were carried out with SPSS. Normality analyses indicated that the data were skewed, leading to the use of non-parametric statistics to answer the research questions. Friedman tests, followed by Post-hoc Bonferroni adjusted Wilcoxon signed-rank tests were employed to examine the intelligibility of the target words and vowels (RQ1-2) and the employment of compensatory strategies (RQ3). To examine the lexical substitutions in more depth, a mixed-methods approach was used where the items causing the most intelligibility breakdowns were identified and tallied together with a qualitative analysis of the patterns observed in the lexical substitutions.

4. Results

This section presents the results concerning the intelligibility and the compensatory strategy analyses.

4.1. Intelligibility

Research questions 1 and 2 sought to examine the intelligibility of the BP EFL learners to the ASR tool. More specifically, we expected to compare the intelligibility of isolated words produced in the present study with paragraph reading data from Kivistö-de Souza and Gottardi (2022) (RQ1) and to extend the results by examining the intelligibility of specific English vowels (RQ2).

Table 3 shows the descriptive statistics for the overall intelligibility data in the present study in comparison to the Kivistö-de Souza and Gottardi (2022) study in regards the BP participants. The present study shows a large amount of variation in the overall intelligibility of the BP speakers, and on average intelligibility was surprisingly low (*M*=55.7%). The low overall intelligibility is also reflected in the amount of intelligibility breakdowns per speaker, ranging from 7 to 21 out of a total of 30 words. Comparing with Kivistö-de Souza and Gottardi's (2022) data, intelligibility in Kivistö-de Souza and Gottardi's (2022) paragraph reading was remarkably higher than in the present study and the number of intelligibility breakdowns per speaker was conversely much lower. Consequently, hypothesis 1 about the intelligibility being higher in isolated words was not confirmed.

Looking at the intelligibility of the target vowels, the high front vowel /i/ was the most intelligible (*M*=73.3%), followed by the low back vowel /ɑ/ (*M*=68.9%). In contrast, the /u/ (*M*=43.3%), /ɛ/ (*M*=45.6%) and /æ/ (*M*=50%) productions were quite unintelligible to the ASR software.

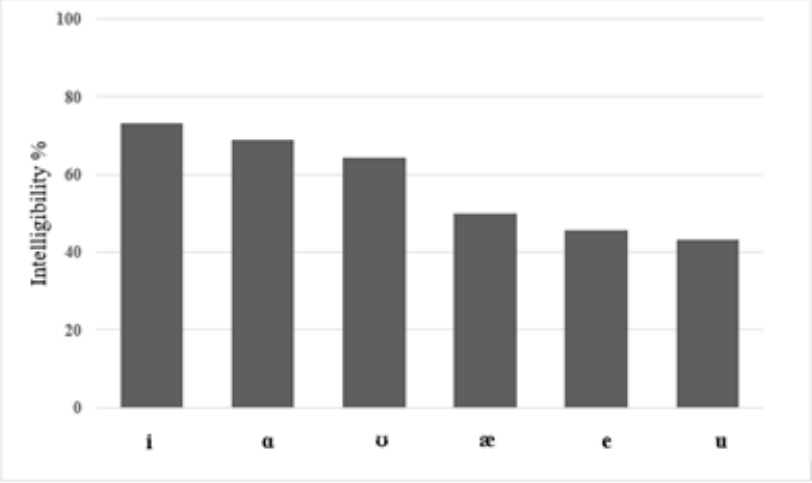
Table 3 - *Intelligibility breakdowns and intelligibility scores averaged across speakers*

	Nº of intelligibility breakdowns averaged across speakers	Intelligibility score %						
		Overall intelligibility	Vowel intelligibility					
			i	ɛ	æ	ɑ	u	ʊ
Present study *N*=18	13.28 (4.21) 7-21	55.74 (14.04) 30-77	73.33 (21.70) 40-100	45.56 (20.36) 20-80	50.00 (25.90) 0-100	68.89 (22.99) 20-100	43.33 (23.00) 20-100	64.44 (26.18) 20-100
Kivistö-de Souza and Gottardi (2022) *N*=15	4.93 (2.93) 1-11	92.80 (4.30) 83-98	-	-	-	-	-	-

Note. Standard deviations between brackets. Range on the final line rounded to whole numbers.

In order to determine whether there were statistical differences in the intelligibility of the target vowels, a Friedman test was carried out with Vowel Type (I /ɛ/ æ/ ɑ/ u/ ʊ) as the independent variable and *Intelligibility* as the dependent variable. Significant difference in the Vowel Type was observed ($\chi 2$ (5) =30.11, $p<.001$), which was further examined with Post-hoc Bonferroni adjusted Wilcoxon signed-rank tests. The post-hoc comparisons further indicated that /ɛ/ and /u/ were significantly less intelligible than /ɑ/ (*p=.043* and p=.032 , respectively), and /i/ (/ɛ-i/ *p=.006,* /u-i/, *p<.005*) The remaining comparisons were not statistically significant. Hypothesis 2 was partly confirmed as intelligibility of individual vowels was low. However, the vowels with the lowest degree of intelligibility were /ɛ/ and /u/ instead of the predicted /æ/ and /ʊ/. The intelligibility of the vowels is visually depicted in Figure 1.

Figure 1 - *Mean intelligibility of the target vowels*



Note. Intelligibility in percentage.

4.2. ASR compensatory strategies

The strategies employed by Microsoft Word Dictate in cases of intelligibility breakdowns were inspected in RQ3. Table 4 indicates the average amount of times a compensatory strategy was employed per speaker. It can be observed that the most frequently used compensatory strategy employed by Microsoft Word Dictate for the isolated words in the present study was substitution, as in Kivistö-de Souza and Gottardi (2022).

Table 4 - *Mean number of compensatory strategies per speaker*

	Present study (*N*=18) Isolated words	Kivistö-de Souza and Gottardi (2022) (*N*=30) Paragraph reading
Substitution	12.28 (4.37)	5.40 (4.91)
Omission	.22 (.55)	0.50 (0.82)
Addition	.39 (.69)	0.07 (0.36)
1x2 Substitution	.39 (.60)	0.20 (0.55)

Note. Standard deviations between brackets.

In order to determine whether there were statistical differences in the use of the compensatory strategies, a Friedman test was carried out with Compensatory Strategy Type (*Substitution/ Omission/ Addition/1x2 Substitution*) as the independent variable and *Compensatory Strategy* as the dependent variable. A significant difference in the Compensatory Strategy Type ($\chi2(3) = 41.43$, $p<.001$) was observed. Post-hoc Bonferroni adjusted Wilcoxon signed-rank tests further showed that *Substitution* significantly differed from the other types ($p<.001$). The remaining comparisons were not statistically significant. Hypothesis 3 was thus confirmed.

To explore the lexical substitutions better, the items that caused the most intelligibility breakdowns were identified. Table 5 shows the most frequent words that caused intelligibility breakdowns and their most frequent substitutions. As can be observed in the table, the majority of words that caused frequent intelligibility breakdowns contained the vowels /ɛ/ and /u/, resulting in 7 of the most 8 frequently mistranscribed items:

Table 5 - *Items with more than 10 intelligibility breakdowns*

Item	Number of intelligibility breakdowns	Frequent lexical substitutions
Pep	17	*Peep* (2); *Back* (2); *But* (2)
Bat	16	*But* (11)
Loot	15	*Look* (9)
Pool	15	*Pull* (5)

Peck	15	*Back* (7)
Luke	13	*Look* (12)
Pet	11	*But* (5)
Full	11	*Four* (5)

Note. In the column frequent lexical substitutions the number between brackets corresponds to the number of occurrences

Here, it can be evidenced that the majority of the most frequent intelligibility breakdowns were substitutions caused by mistranscriptions at the segment level. It is possible to further categorize these breakdowns into two: errors caused by word frequency *(Loot, Pep, Luke, Peck*) and errors caused by high functional load vowels (*Pet, Bat*). The remaining words (*Pool, Full*) in particular had a number of frequent mistranscriptions caused by various neighboring minimal pairs or near minimal pair words (*Cool, Pull, Poop, Full, Four*).

Thus, it can be assumed that the primary issues leading to intelligibility breakdowns were either vocalic (unintelligible vowel production), phonetic or syntactic- either due to missing contextual clues or due to word frequency issues. For example, the consistent mistranscriptions of the words *Look* and *Pool* as *Luck* and *Pull* are likely due to participants producing a too short vowel, which was interpreted as /ʊ/ instead of /u/. This exemplifies how BP speakers may not necessarily acknowledge the two contrastive English vowels as two different sound categories, assimilating both into one singular L1-FL composite category. It is also possible that participants' aspiration in some items was insufficient as several substitutions involved the interpretation of a voiceless plosive as a voiced one: *Pep* as *Back*, *Peck* as *Back* and *Pet* as *But*. Regarding the words *Pet* and *Bat,* it is interesting to denote that the majority of mistranscriptions resulted in the word *but*, which resembles the findings in Kivistö-de Souza and Gottardi (2022) regarding the intelligibility breakdowns with the word *Bob*. It is possible that the ASR tool did not have enough contextual clues to understand which word to prioritize in the transcription, defaulting to most common CVC word transcriptions between two plosives (*But; That)*.

5. Discussion

The present study set out to examine how Microsoft Word Dictate assesses the intelligibility of BP speakers of English in individual words containing tricky EFL vowels. Overall, the results showed that the intelligibility was lower than expected and that the vowels /ɛ- æ/ and /u- ʊ/ posed significant intelligibility challenges. As in Kivistö-de Souza and Gottardi (2022), the ASR software's preferred strategy when encountering an intelligibility breakdown was to substitute the intended word with another lexical item.

The overall low intelligibility rate in the present study was a somewhat surprising finding and did not let us confirm that employing words in isolation would result in more accurate transcriptions than presenting continuous speech,

such as the paragraph reading task in Kivistö-de Souza and Gottardi (2022), contrary to our initial hypothesis. Whereas Kivistö-de Souza and Gottardi (2022) noted that connected speech phenomena such as linking, vowel reduction and speech rate affected intelligibility negatively, the results of the present study do not seem to favor isolated word pronunciation practice either. Nevertheless, caution should be used when interpreting these results as participants and the methodology in the two studies were different. On the other hand, it might be that the contextual cues present in continuous speech make the completion of the intended output easier for the ASR tools when the accuracy of the speech signal is not optimal. The present study prioritized phonetic environment in the creation of the target words, which meant that some infrequent items (*Pep, Luke*) were included. It can be assumed that infrequent words presented in isolated form leave very little contextual clues for Microsoft Word Dictate to "guess" the correct transcription, especially since ASR tools are built on speech data and vocabulary corpuses. However, Kivistö-de Souza and Gottardi (2022) also reported a number of infrequent and unfamiliar words present in their paragraph reading approach, which suggests the presence of additional words in a sentence may contribute significantly for intelligibility scores in Microsoft Word Dictate.

While at first sight it may appear that the low intelligibility scores observed in the present study indicate that L2 users desiring to use ASR tools to develop their EFL segmental production should not employ isolated words, we would like to put forward a different viewpoint. The transcriptions that do not match the intended message, as observed in the present study may contribute to learners' noticing of pronunciation issues in their production. That is, as an auxiliary tool, ASR can aid learners in becoming more aware of their pronunciation, thus possibly aiding the development of the target sound categories. Therefore, despite paragraph reading tasks proving higher intelligibility values, using ASR for isolated words may still play a significant role in contributing with the noticing of specific phonological issues at the segmental level.

The inaccurate transcriptions for the vowel pairs /ɛ- æ/ and /u- ʊ/ corroborate previous research regarding the difficulty of English vowel production for BP speakers (Lima Junior., 2017). However, in the present study, the vowels that were among the least intelligible (/ɛ/ and /u/) were those that phonetically should not cause problems for Brazilian Portuguese speakers. Participants' productions were not presented to human judges or acoustically assessed, so we cannot entirely disregard the idea that the participants in the present study produced these vowels inaccurately. We do not find this plausible though and suggest that other factors, than the coding of the actual speech signal are taking place when ASR software transcribes speech. One of these could be the functional load. Namely, when a word has many minimal pairs (i.e., it has a high functional load), the chances for mistranscriptions are higher when a vowel is not accurately pronounced (*Pet* was transcribed as a number of similar words such as *That, Pat, Bet,* and *Bad).* Another problematic factor previously mentioned includes participants' productions of unfamiliar words present in this research. By the same degree that

the ASR tool is likely to face difficulties in transcribing words such as *Pep* and *Luke*, it is also plausible that the participants would also face equal challenge in producing these words intelligibly. In particular, since many of the problematic words included the target vowels /ɛ/ and /u/, this could have had a significant impact on the vowel intelligibility scores, and thus have resulted in the unusual findings regarding which vowels BP speakers present in this study had most intelligibility breakdowns.

The findings on compensatory strategies confirmed our hypothesis by following a similar pattern found in Kivistö-de Souza and Gottardi (2022). *Substitution* was the dominant compensatory strategy used by Microsoft Word Dictate when transcribing BP speakers of English, and there was a significantly smaller number of other compensatory strategies. However, the disparity of strategies used by Microsoft Word Dictate was larger in the present study. It could be argued that this effect on the compensatory strategies was due to the nature of the speech being transcribed (segments). In particular, it can be assumed that Microsoft Word Dictate had less contextual clues in order to perform compensatory strategies that include *Addition* and *1x2 Substitutions*, and the nature of the test performed could have significantly lowered the occurrence of *Omissions*, as Microsoft Word Dictate had ample time to transcribe the single word input that was being uttered by participants. Furthermore, this could also explain why additions occurred more frequently, as Microsoft Word Dictate continuously attempted to transcribe as if it was beginning a sentence (e.g.: *sad* transcribed as *It's sad*; pet transcribed as *The pet*)

6. Limitations and Conclusions

Due to the inconclusive results in the present study in relation to the employment of individual words vs. connected speech, future studies could employ a method of embedding target words into carrier sentences, e.g. 'I say ___ again'. This method would allow the use of specific target sounds in a highly controlled environment, but would also offer Microsoft Word Dictate a 'context', which might reduce the occurrence of adding extra words to the transcriptions. Nevertheless, we should state that in the present study, a specific ASR tool was employed and the results observed here are only applicable to Microsoft Word Dictate.

Other limitations of the study were not having assessed the participants' actual pronunciation accuracy through subjective (human) or objective measures, which would have aided to answer the question of whether the problems in intelligibility were due to participants' mispronunciations or to faults in the ASR software. Furthermore, EFL proficiency was not controlled in the present study as the EFL proficiency measures were based on participants' self-reports. Future studies should also consider measuring participants' affective factors, such as FL anxiety, pronunciation anxiety and motivation, as these factors might play a role in who benefits from the use of ASR technology in autonomous EFL pronunciation

development. Future studies should also look at the learners' opinions on the use of ASR, as even the best tool is unlikely to be used if the learners dislike using it. Future studies should also measure what learners do with the ASR output: if learners do not understand that the output differs from the intended message because of a pronunciation problem, the use of the software for pronunciation development purposes is pointless.

Ultimately, we still believe in the relevance of Microsoft Word Dictate (and ASR as a whole) in providing a safe environment that can aid EFL learners in becoming aware of possible intelligibility problems in their own productions, allowing instructors to find additional ways of providing for more input and more time for EFL pronunciation learning outside of the classroom context. More specifically, even though not measured in the present study, we hypothesize that tasks that focus on the production of isolated words are likely to make the intelligibility problems with specific segments more visually salient to the learner than when these are embedded in longer stretches of speech (e.g. paragraph or sentence reading). Learners are nevertheless unlikely to be able to develop their pronunciation entirely autonomously with the output provided by ASR tools, and instructor's guidance of how to make the most of this technology is thus necessary. In this way, instructors can make use of ASR technology as a bridge between traditional language mediums and the novel machine learning language models that are making their way into the EFL classrooms.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001, granted to the first author.

Note

1. The research project was approved under the CAEE 67447723.6.0000.0121 by the Ethics Committee at the authors' institution.

References

Baran-Łucarz, M. (2014). The link between pronunciation anxiety and willingness to communicate in the foreign-language classroom: The Polish EFL context, *Canadian Modern Language Review-Revue Canadienne Des Langues Vivantes, 70, n. 4,* 445-473.

Baran-Łucarz, M. (2016). Conceptualizing and measuring the construct of pronunciation anxiety: results of a pilot study, M. Pawlak, *Classroom-Oriented Research: Reconciling Theory and Practice,* 39-56. Springer International Publishing. https://doi.org/10.1007/978-3-319-30373-4_3

Bashori, M., van Hout, R., Strik, H., & Cucchiarini, C. (2021). Effects of ASR-based websites on EFL learners' vocabulary, speaking anxiety, and language enjoyment, *System (Linköping)*, *99*, 102496. https://doi.org/10.1016/j.system.2021.102496

Bion, R., Escudero, P., Rauber, A., & Baptista, B. (2006). Category formation and the role of spectral quality in the perception and production of English front vowels,

Interspeech 2006 - *ICSLP, Ninth international conference on spoken language processing,* Pittsburg. https://doi.org/10.21437/Interspeech.2006-411

Burleson, D. (2007). Improving intelligibility of non-native speech with computer-assisted phonological training, *IULC Working Papers Online.*

Celce-Murcia, M.; Brinton, D., & Goodwin, J. (2010). *Teaching pronunciation: A course book and reference guide*, New York: Cambridge University Press.

Chun, D. M. (2011). Computer-assisted language learning, *Handbook of Research in Second Language Teaching and Learning.* New York: Routledge, 663-680.

Chun, D. M. (2013). Computer-assisted pronunciation teaching, The Encyclopedia of Applied Linguistics. New Jersey: Blackwell Publishing Ltd. 1-11

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

Derwing, T. (2010). Utopian goals for pronunciation teaching, *Pronunciation in second language learning and teaching proceedings,* Iowa: Iowa State University.

Derwing, T. M., & Munro, M. J. (2022). Pronunciation learning and teaching, Derwing, T. M., Munro, M. J., & Thomson, R. I. *The Routledge handbook of second language acquisition and speaking.* New York: Routledge, 147-159.

Derwing, T. M., Munro, M. J., & Carbonaro, M. (2000). Does popular speech recognition software work with ESL speech? *Tesol Quarterly*, *34*(3), 592-603. https://doi.org/10.2307/3587748

Flege, J., & Bohn, O.-S. (2021). The Revised Speech Learning model (SLM-r). *Second Language Speech Learning: Theoretical and Empirical Progress*, Cambridge: Cambridge University Press. 3-83.

Gonçalves, A. R., & Silveira, R. (2015). Frequency effects on the intelligibility of English words with high front vowels, *Organon,* 30(58). 127-152

Hincks, R. (2003). Speech technologies for pronunciation feedback and evaluation, *ReCALL, 15*, 3-20. https://doi.org/10.1017/S0958344003000211

Jin, S.H., & Liu, C. (2014). Intelligibility of American English vowels and consonants spoken by international students in the US, *Journal of speech, language, and hearing research : JSLHR,* 57. https://doi.org/10.1044/2014_JSLHR-H-13-0047

Juang, B., & Rabiner, L. (2005). Automatic Speech Recognition - a brief history of the technology development, *Elsevier Encyclopedia of Language and Linguistics.*

Kang, O., Thomson, R. I.; & Moran, M. (2018). Empirical approaches to measuring the intelligibility of different varieties of English in predicting listener comprehension, *Language Learning, 68*(1), 115-146.

Kivistö-de Souza, H., Gottardi, W. (2022). How well can ASR technology understand foreign-accented speech? *Trabalhos em Linguística Aplicada*, *61*. 10.1590/010318138668782v61n32022.

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

Levis, J. M., & Silpachai, A. O. (2022). Speech intelligibility, Derwing, T. M., Munro, M. J., & Thomson, R. I. *The Routledge Handbook of Second Language Acquisition and Speaking.* New York: Routledge, *10*, 160-173.

Levis, J. M., & Suvorov, R. (2012). Automatic Speech Recognition, *Encyclopedia of Applied Linguistics* 10.1002/9781405198431.wbeal0066.

Lima Junior, R. M. (2015). A influência da idade na aquisição de seis vogais do inglês por alunos brasileiros, *Organon, 30*.

Lima Junior, R. M. (2017). The influence of metalinguistic knowledge of segmental phonology on the production of English vowels by Brazilian undergraduate students, *Ilha do Desterro*, *70,* 17-130,

Lima Junior., R. M. (2019). A longitudinal study on the acquisition of six English vowels by Brazilian learners, *Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS)*, Melbourne. Canberra: Australasian Speech Science and Technology Association Inc. 3180-3184.

McCrocklin, S., & Edalatishams, I. (2020). Revisiting popular speech recognition software for ESL speech, *Tesol Quarterly*, *54(4),* 1086-1097. https://doi.org/10.1002/tesq.3006

Miller, N. (2013). Measuring up to speech intelligibility, *International Journal of Language & Communication Disorders*, *48*(6), 601-612. https://doi.org/https://doi.org/10.1111/1460-6984.12061

Mroz, A. (2018). Seeing how people hear you: French learners experiencing intelligibility through automatic speech recognition, *Foreign Language Annals*, *51*. https://doi.org/10.1111/flan.12348

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners, *Language Learning*, *45*(1), 73-97. https://doi.org/https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Munro, M. J., & Derwing, T. M. (1999). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners, *Language Learning*, *49*(Suppl 1), 285-310. https://doi.org/10.1111/0023-8333.49.s1.8

Munro, M. J., Derwing, T. M., & Morton, S. L. (2006). The mutual intelligibility of L2 speech, *Studies in Second Language Acquisition*, *28*(1), 111-131. https://doi.org/10.1017/S0272263106060049

Neri, A., Cucchiarini, C., & Strik, H. (2008). The effectiveness of computer-based speech corrective feedback for improving segmental quality in L2 Dutch, *ReCALL © European Association for Computer Assisted Language Learning*, *20*, 225-243. https://doi.org/10.1017/S0958344008000724

Nobre-Oliveira, D. (2007). *The effect of perceptual training on the learning of English vowels by Brazilian Portuguese speakers*, [Doctoral dissertation - Universidade Federal de Santa Catarina]

Phillips, E. M. (1992). The effects of language anxiety on students' oral test performance and attitudes, *Modern Language Journal, 76*(1), 14–26. https://doi.org/10.2307/329894

Price, M. L. (1991). The Subjective Experience of Foreign Language Anxiety: Interview with Highly Anxious Students. E. K. Horwitz, & D. J. Young, L*anguage Anxiety: From Theory and Research to Classroom Implications,* 101-108. Englewood Cliffs, NJ: Prentice Hall.

Rauber, A. S. (2006). *Perception and production of English vowels by Brazilian EEFL speakers,* [Doctoral dissertation - Universidade Federal de Santa Catarina]

Shibano, T., Zhang, X., Li, M., Cho, H., Sullivan, P., & Abdul-Mageed, M. (2021). Speech technology for everyone: Automatic Speech Recognition for non-native English, *ArXiv*, *abs/2110.00678*.

Silveira, R. (2004). *The influence of pronunciation instruction on the perception and the production of English word-final consonants,* [Doctoral dissertation - Universidade Federal de Santa Catarina]

Silveira, R., Zimmer, M., Alves, U. K.(2009). *Pronunciation instruction for Brazilians: bringing theory and practice together*, Newcastle: Cambridge Scholars Publishing,

Tejedor-García, C., Escudero-Mancebo, D., Cámara-Arenas, E., González-Ferreras, C., & Cardeñoso-Payo, V. (2020). Assessing pronunciation improvement in students of English Using a controlled computer-assisted pronunciation tool, *IEEE Transactions on Learning Technologies*, *13*(2), 269-282. https://doi.org/10.1109/TLT.2020.2980261

Thomson, R., & Derwing, T. (2014). The effectiveness of L2 pronunciation instruction: A Narrative Review, *Applied Linguistics*, *2014*, 1-20. https://doi.org/10.1093/applin/amu076

Wachowicz, K. A., & Scott, B. (1999). Software that listens: It's not a question of whether, it's a question of how, *CALICO Journal*, *16*(3), 253-276.

Weinberger, S. (2015). Speech Accent Archive. George Mason University, Retrieved from http://accent.gmu.edu

Young, D.J. (1992). Language Anxiety from the Foreign Language Specialist's Perspective: Interviews with Krashen, Omaggio Hadley, Terrell, and Rardin, *Foreign Language Annals, 25:* 157-172. https://doi.org/10.1111/j.1944-9720.1992.tb00524.x

Recebido em: 04/03/2024
Aceito em: 15/10/2024