Automatic speech recognition and text-to-speech technologies for L2 pronunciation improvement: reflections on their affordances

Tecnologias de reconhecimento automático da fala e texto-fala para o aprimoramento da pronúncia em L2: reflexões das suas aplicabilidades

William Gottardi *[1], **Janaina Fernanda de Almeida** †[1] and **Celso Henrique Soufen Tumolo** ‡[1]

[1] *Universidade Federal de Santa Catarina, Centro de Comunicação e Expressão, Florianópolis, SC, Brasil.*

Abstract

This paper presents a reflection on two technologies - automatic speech recognition (ASR) and Text-to-Speech (TTS) - to improve learners' pronunciation, aiming for successful spoken communication. It sheds some light on the practical usage of these technologies, demonstrating their effectiveness, qualities, and limitations to assist teachers in deciding the most efficient digital resources applied to their students' needs. A review of literature on previous empirical studies was carried out, with quantitative and/or qualitative studies conducted by researchers in the field, investigating teachers' and learners' perceptions and the use of ASR and TTS as a pedagogical tool for pronunciation practice. As a result, it was concluded that a) the presented resources seem to have the potential to enhance pronunciation practice, both in terms of perception and production; b) technology can result in considerable benefits to learners, mainly as a supplement to pronunciation teaching; and c) the use of these digital resources is a way of giving learners the opportunity to focus on their specific difficulties and receive personalized feedback while becoming more autonomous in their learning process.

Keywords: Automatic speech recognition. Text-to-speech. CALL. Pronunciation teaching. Pronunciation improvement.

Resumo

Este artigo apresenta uma reflexão sobre duas tecnologias – reconhecimento automático da fala (ASR – Automatic Speech Recognition) e texto-fala (TTS – Text-to-Speech) – para aprimorar a pronúncia dos alunos, visando a uma comunicação oral competente. O trabalho explora o uso dessas tecnologias, demonstrando sua eficácia, qualidades e limitações para ajudar os professores a decidirem os recursos digitais mais eficientes aplicados às necessidades de seus alunos. Foi realizada uma revisão bibliográfica de estudos empíricos prévios, com pesquisas quantitativas e/ou qualitativas realizadas por pesquisadores da área, investigando as percepções de professores e alunos e o uso de ASR e TTS como ferramentas pedagógicas para o ensino de pronúncia. Como resultado, concluiu-se que a) os recursos apresentados demonstram ter potencial para aprimorar a prática da pronúncia, tanto em termos de percepção como produção; b) a tecnologia pode resultar em benefícios consideráveis para os alunos, principalmente como um suplemento ao ensino de pronúncia; e c) o uso desses recursos digitais é uma forma de dar aos alunos a oportunidade de focar em suas dificuldades específicas e receber um retorno personalizado, tornando-os mais autônomos em seu processo de aprendizagem.

Palavras-chave: Reconhecimento automático da fala. Texto-fala. CALL. Ensino de pronúncia. Aprimoramento de pronúncia.

Textolivre
Linguagem e Tecnologia

DOI: 10.35699/1983-3652.2022.36736

Session:
Articles

Corresponding author:
William Gottardi

Section Editor:
Danniervelin Pereira
Layout editor:
Danniervelin Pereira

Received on:
October 12, 2021
Accepted on:
December 2, 2021
Published on:
February 10, 2022

1 Introduction

Spoken communication to be effective must consider pronunciation as a fundamental part of it (PENNINGTON; ROGERSON-REVELL, 2019) since the primary medium of language is speech

*Email: william.gottardi@posgrad.ufsc.br
†Email: janainafernandadealmeida@gmail.com
‡Email: celsotumolo@yahoo.com.br

(SLABAKOVA, 2016). Therefore, when it comes to second language acquisition (SLA), we must consider the different dimensions of second language (L2) speech: intelligibility, comprehensibility, and accentedness. Concerning intelligibility, Munro and Derwing (1995, p. 289) broadly define it as "the extent to which a speaker's message is actually understood by a listener, but there is no universally accepted way of assessing it". It differs from comprehensibility – the listener's evaluation of difficulty in understanding another person's speech, and from accentedness – the distinction of the pronunciation of a sentence sounds in comparison to an expected production pattern (MUNRO; DERWING; MORTON, 2006).

For some authors, the goal of L2 pronunciation teaching and research should be enhanced intelligibility and comprehensibility instead of native-likeness (O'BRIEN et al., 2018), since "rather than requiring native-sounding oral output, L2 users need intelligible speech" (MUNOZ, 2008, p. 213). Moreover, pronunciation difficulties in an additional language can compromise intelligibility, which may also hinder comprehension, compromising oral communication (SICOLA; DARCY, 2015). Although pronunciation plays a critical role in successful communication, pronunciation teaching is sometimes neglected due to time constraints and insecurity regarding how to get started (ROCCAMO, 2014). Furthermore, pronunciation is connected to identity issues and language attitudes, for both learners and teachers, which might impact teachers' confidence and willingness to teach pronunciation (PENNINGTON; ROGERSON-REVELL, 2019).

In order to strengthen the potential benefits of pronunciation teaching, learners' attention should be directed to those aspects that are likely to most affect their oral performance. According to Derwing (2018), it is only possible to value the effectiveness of pronunciation teaching if it improves communication in general; that is, once the pedagogical intervention helps to increase learners' intelligibility and comprehensibility. Therefore, the objective of this paper is to present a reflection on two technologies - automatic speech recognition (ASR) and text-to-speech (TTS) - to improve learners' pronunciation, aiming for successful spoken communication. For this purpose, the reflection will address the affordances of the technologies by reviewing both quantitative and qualitative pieces of research from authors in the field.

To start with, the following section will detail recent empirical findings on pronunciation teaching, especially as a means to draw learners' attention to L2 phonological forms of the target language input. By focusing on these aspects, learners can achieve improvements in their L2 pronunciation skills in order to reduce communication breakdowns due to inaccuracy in speech perception or production and, in turn, enhance speech intelligibility and comprehensibility.

2 Pronunciation instruction

In the past few years, several studies focusing on L2 pronunciation teaching have been carried out, reporting benefits for the acquisition of both segmental and suprasegmental features[1] (THOMSON; DERWING, 2014). This section aims to consider recent claims on the discussion of pronunciation teaching, defending it as a means to help L2 learners overcome difficulties in their pronunciation skills that otherwise could affect overall communication in the target language. Although there is also a great deal of current discussion concerning which pronunciation feature yields more significant pronunciation gains, the comparison of the specific results for suprasegmental or segmental instruction is beyond the scope of this study (for a review, see (GORDON; DARCY, 2016; LEE; PLONSKY; SAITO, 2020; ZHANG; YUAN, 2020)). However, a significant finding from previous research is that pronunciation instruction tends to lead to more improvement gains when employing explicit techniques rather than implicit ones (THOMSON; DERWING, 2014; GORDON; DARCY, 2016).

Such a finding endorses discussions on the role of oral input for L2 learning. In general terms, input is the language embedded in the communication contexts to which learners are exposed during their learning process (VANPATTEN; SMITH; BENATI, 2019). Accordingly, it is common knowledge that oral input is crucial for acquiring the target language phonology (TYLER, 2019). More specifically,

1 Segmental features are the phonetic features at the segment level, distinguishing the sounds of a given language (e.g., vowels and consonants). In contrast, suprasegmental features are those beyond the level of individual sounds, such as stress and intonation (YAVAŞ, 2011).

input is a prerequisite for the development of learners' perceptual skills as it provides the phonological data for one to perceive the L2 sounds (LIAKIN; CARDOSO; LIAKINA, 2017; TYLER, 2019).

However, when it comes to instructional language learning settings, language input is somehow limited in terms of quality or/and quantity. As observed by Munoz (2008), the exposure to the target language is restricted to short class sessions, and it comes primarily from the teacher and peers, who usually share the same L1. Therefore, considering the lack of frequent language input and the reduced opportunities for practicing oral skills (CARLET; KIVISTÖ-DE SOUZA, 2018), L2 learners may not be able to fully develop their L2 pronunciation skills only through the limited contact with the target language provided in class.

In order to ensure learning, exposure to the target language has to provide access to comprehensible input, that is, learners need to make sense of the language instances that are being presented. According to VanPatten, Smith, and Benati (2019, p. 46), the term comprehensible input is associated with Krashen's ideas (proposed in the late 1970s), and it is based on the assumption that "during the act of comprehension, learners are engaged in mapping meaning onto form". Therefore, on one hand, input is only effective for L2 acquisition if the learner can comprehend it, or else the internal mechanism cannot use the presented data to extract its meaning. On the other hand, as learners process input primarily for meaning (VANPATTEN, 2008), their focal attention on the phonological aspects of the language is likely to be diminished.

Under these circumstances, pronunciation teaching appears to be crucial in overcoming the shortcomings of classroom settings. Following Thomson and Derwing (2018, p. 340)' conclusion, explicit teaching is likely to have a positive impact on the acquisition of phonological forms because "it orients learners' attention to phonetic information, which promotes learning in a way that naturalistic input does not". Consequently, learners who are more aware of the underlying phonological forms are inclined to achieve a more target-like performance at both perception and production levels (CARLET; KIVISTÖ-DE SOUZA, 2018).

According to Carlet and Kivistö-de Souza (2018, p. 104), L2 phonological awareness "can be developed through any activity that brings a specific aspect into the language learners' consciousness". The authors also provide some examples of consciousness-raising activities reported in the literature, such as the explicit comparison between the L1-target language phonologies, input enhancement, and feedback techniques. Ultimately, the authors defend that helping learners to raise their awareness of the target language phonology "does not only positively reflect on their L2 pronunciation, but also enables them to take control of their pronunciation learning by developing self-monitoring abilities" (CARLET; KIVISTÖ-DE SOUZA, 2018, p. 104).

In a similar vein, Darcy (2018) stresses that feedback is also a predictor of self-awareness pronunciation development, mainly because it indicates specific difficulties to the learner as they occur. However, due to the different types of feedback, the author points out that explicit feedback should be favored when pronunciation aspects are taught as an integrated part of a lesson. Therefore, explicit feedback helps to clarify that the correction is about a particular form rather than meaning. In this manner, it is possible to draw the learners' attention to their production compared to what they were expected to produce. In response, learners can focus on monitoring their pronunciation in order to achieve a more intelligible and comprehensible speech.

Learners' production also plays a pivotal role in pronunciation learning since output is crucial for this skill improvement (DEMENKO; WAGNER; CYLWIK, 2010). Output also has the function of promoting automaticity, freeing limited cognitive resources (e.g., working memory, and attention), and letting them available to other language acquisition processes (GRASS; MACKEY, 2015; ORTEGA, 2009). In addition, like any other skill, practice leads to proceduralized knowledge and, after consistent practice, procedural knowledge becomes automatic knowledge, which facilitates a fluent and spontaneous speech (DEKEYSER, 2015).

Considering the discussion above, pronunciation aspects should be explicitly taught as early as possible in the learning process (DARCY, 2018; DERWING, 2018), along with activities aiming at developing learners' awareness of the specific forms and their own oral performance. Also, the efficiency of pronunciation teaching should be built on the three main 'ingredients' of explicit and communicative

activities - containing or not repetition (focalizing on both form and meaning) - of focus on perception, and of explicit feedback (DARCY, 2018). Thus, in agreement with Thomson and Derwing (2014) that informed instruction combined with practice opportunities will help learners improve speech production and considering the importance of pronunciation to spoken communication and pronunciation teaching to SLA, all the resources available to help accomplish the goal of pronunciation development are welcome. In this matter, digital resources can be beneficial to teachers and learners, as explored in the next section.

3 Technology and pronunciation teaching

Language classrooms without any form of technology would create a limited and artificial learning environment once technology has been so interwoven and pervasive in human activities (CHUN; KERN; SMITH, 2016). However, the usage of technology during the lessons does not make inefficient pedagogy efficient (GOLONKA et al., 2014).

Considering the importance of reflecting on the usage of digital technology for L2, there is a specific subfield of Applied Linguistics that studies the relationship between technology and SLA called Computer-Assisted Language Learning (CALL) (MARTINS; MOREIRA, 2012). Davies (2006, p. 261) defines CALL as "an approach to language teaching and learning in which computer technology is used as an aid to the presentation, reinforcement, and assessment of material to be learned, usually including a substantial interactive element.". This field has grown quickly in recent decades (PENNINGTON; ROGERSON-REVELL, 2019) and it comprehends a wide array of practices, which predicts equally varied outcomes and pedagogical effectiveness (LEVIS; SUVOROV, 2013a). This tension between technology and pedagogy is a key issue concerning the topic (ROGERSON-REVELL, 2021). For this reason, we now turn our focus on the two technologies with supporting evidence for L2 pronunciation improvement along with some classroom implications. In section 3.1, we explore automatic speech recognition (ASR) technology as an additional resource for learners to produce more oral output with explicit feedback; and in section 3.2, we present text-to-speech (TTS) technology and its affordances for pronunciation improvement, focusing on input, that is, "the sine qua non of acquisition" (GRASS; MACKEY, 2015, p. 177).

3.1 Automatic speech recognition and its affordances for pronunciation improvement

ASR can be defined as "an independent, machine-based process of decoding and transcribing oral speech" (LEVIS; SUVOROV, 2013b, p. 316). By building a string of words from an acoustic signal, it can be applied to dictation (a single specific speaker's monologue transcription) tools, or human-computer interaction (JURAFSKY; MARTIN, 2000), such as Intelligent Personal Assistant (IPA) embedded to smart devices (e.g., smartphones and smart speakers) (INCEOGLU; LIM; CHEN, 2020; MOUSSALLI; CARDOSO, 2020). This technology started to be developed by the late 1940s and early 1950s. However, it has been constantly enhanced due to the development of new model techniques and algorithms, improvement in noisy speech recognition, and the demand to integrate it into mobile devices (JURAFSKY; MARTIN, 2000; LEVIS; SUVOROV, 2013b).

ASR technology applied to a CALL context has been criticized by some authors in the past decade due to its incapacity to comprehend L2 speech accurately at a similar rate as human listeners (DERWING; MUNRO; CARBONARO, 2000; KIM, 2006; LEVIS; SUVOROV, 2013b), its much lower accuracy scores for nonnative speakers than those for native speakers (ASHWELL; ELAM, 2017; ROGERSON-REVELL, 2021), and its insufficient or even incorrect feedback (CHEN, 2011; DEMENKO; WAGNER; CYLWIK, 2010; LEVIS; SUVOROV, 2013b; ROGERSON-REVELL, 2021). On the other hand, recent research has shown that this technology has been improving in the past years (ASHWELL; ELAM, 2017; DIZON, 2020; DIZON; TANG, 2020; MCCROCKLIN; EDALATISHAMS, 2020; MOUSSALLI; CARDOSO, 2020; BOGACH et al., 2021). Furthermore, Ashwell and Elam (2017, p. 61) argue that "these systems are continually improving on their respective accuracy rates by constantly gathering acoustic information and utilizing machine learning".

ASR systems usually use a native speaker as a model from a database containing a vast number of native speaker speech samples. Although it has been advancing in recent years, especially considering

native speaker recognition, the accuracy level for nonnative speech is considerably lower (REVELL-ROGERSON, 2021). Notwithstanding, in a recent study exploring the accuracy of Google Voice Typing, a free ASR-based dictation tool, considering native and nonnative English speakers' samples, McCrocklin and Edalatishams (2020, p. 1092) concluded that "across all L2 speech samples, there was a statistically significant relationship between Google recognition and human listener's intelligibility as well as ratings of comprehensibility". The authors calculated descriptive statistics over 60 sentences produced by the participants of the study. Each speaker dictated each sentence twice to a Google Document. The participants were divided into three groups according to their L1 – English L1 (n = 10), Spanish L1 (n = 10), and Mandarin Chinese L1 (n = 10). The results show an accuracy rate of 96.2%, 92.7%, and 90.9, respectively. The authors state that "whereas earlier research found recognition of nonnative speech 18–20% lower than native speech, Google has reduced that gap to 3–5%." (MCCROCKLIN; EDALATISHAMS, 2020, p. 1094).

In a different study, Ashwell and Elam (2017) investigated how accurately Google Web Speech API could recognize the speech of Japanese learners of English as a foreign language (EFL). Participants produced 13 sentences containing specific grammatical features as an elicited imitation test. They found that the system had an overall recognition accuracy of 89.4%. They concluded that the pronunciation of specific sounds is the most problematic issue for the systems to perform the speech recognition process if compared to native speaker input. In addition, they affirm that pronunciation issues may not be a barrier for ASR systems and this technology could be used for assessing student's grammatical ability.

Following a pedagogical point of view, Inceoglu, Lim, and Chen (2020) explored the usefulness of ASR pronunciation practice to check its effects on learner's production in a segmental level as well as the learners' perception of the usage of ASR as a learning tool. A total of 19 Korean university students produced 28 minimal pair sentences containing vowel contrasts in a pretest and posttest study design. Results of acoustic analysis showed a meaningful improvement in some vowels, but no changes in others. However, the great majority of the participants indicated that ASR is useful for pronunciation practice.

Considering learner's perception, Mroz (2018) investigated how 16 learners of French as a foreign language using ASR in Gmail developed greater awareness of their own intelligibility. The author followed a qualitative approach analyzing participants' responses to semistructured interviews. The results show that most participants considered ASR to be a relevant diagnostic tool once they could assess the gaps and successes in their intelligibility by using such technology.

Moreover, Golonka et al. (2014) reviewed over 350 studies to comprehend the learning and teaching effectiveness of technology use. The authors reviewed empirical studies that compared the use of newer technologies to traditional materials and methods. The review considered individual study tools (e.g., grammar checker, ASR, electronic dictionary, and pronunciation programs), classroom-based technologies, mobile devices, and network-based social computing. They found strong support for a positive impact of ASR programs on foreign language (FL) learning and teaching. In addition, they affirmed that "ASR technology can facilitate improvement in pronunciation to a larger extent than human teachers can and [...] ASR programs have great potential in FL learning" Golonka et al. (2014, p. 88).

Furthermore, ASR can be applied in varied ways to facilitate the learning process, inside and outside the classroom (KIM, 2006; CHEN, 2011; LEVIS; SUVOROV, 2013b; GOLONKA et al., 2014; ASHWELL; ELAM, 2017; LIAKIN; CARDOSO; LIAKINA, 2017; MROZ, 2018; DIZON; TANG, 2020; DIZON, 2020; INCEOGLU; LIM; CHEN, 2020; MCCROCKLIN; EDALATISHAMS, 2020; ROGERSON-REVELL, 2021). To put it concisely, based on the aforementioned studies, ASR tools can contribute to pronunciation improvement by:

- allowing the development of L2 learners' autonomy, offering an opportunity for learners to work on their pronunciation individually, at a self-selected pace;
- encouraging learners to produce more output in a low-anxiety environment, talking to a tireless listener (the algorithm);
- helping learners to improve not only their pronunciation but also their oral communication skills,

speaking fluency, and accuracy;

- increasing learners' confidence and motivation, by engaging students in the process of learning and fostering a more positive attitude towards it;
- providing learners with the opportunity to receive pronunciation feedback outside the language classroom (from the application);
- enabling ubiquitous, out-of-class learning, which allows learners to decide when, what, and how to learn;
- enabling learners to interact with IPAs (e.g.: Siri, Microsoft Cortana, or Google Assistant), performing spontaneous, meaningful, and authentic communicative tasks, also offering an opportunity to test their ability to produce intelligible speech; and
- facilitating the extensive practice of segmental and suprasegmental features of the language, from minimal pair to mirroring famous speeches or rehearsing presentations.

Besides all these contributions, ASR offers limitless opportunities to practice oral output. As mentioned in section 2, output plays a pivotal role in second language acquisition processes. ASR is, therefore, a versatile resource that can be used to transcribe an audio file to text, automatically generate subtitles in a video on YouTube, interact with a smart device using voice commands, practice pronunciation autonomously through mobile applications and use dictation programs. Moreover, it is responsive to different learning goals and adaptable to any language curriculum. As Yoshida (2018) points out, ASR has become available in different programs as built-in features. Many programs are available for free on the internet (e.g.: Google Docs' voice typing, Microsoft Word's dictation tool, or Google and Bing's voice search). Some of them do not even require an internet connection (speech recognition in Windows 10, for instance).

This extended practice that ASR places at disposal is "especially significant for learners who have little to no access to other L2 speakers outside of class" (DIZON; TANG, 2020, p. 108) and hence with less opportunities to practice their oral skills in general. By practicing with the ASR program, the listener will be the algorithm that transcribes the learner's utterances. Thus, learners can check to which extent the application could understand their speech and keep practicing until the software transcribes the intended utterance correctly, that is, their speech was intelligible to the application and the communicative purpose was fulfilled. Although ASR is an additional tool for learning and does not substitute social interaction, this oral practice is of paramount importance once "many individuals in EFL settings have a strong need to improve their oral abilities" (CHEN, 2011, p. 60).

Furthermore, the overall teachers and students' perceptions of ASR applied to teaching and learning demonstrate a bright future for this technology. Levis and Suvorov (2013b, p. 319) state that many studies indicate that "software that includes ASR is a huge plus to language learners in terms of practice, motivation, and the feeling that they are actually communicating in the language rather than simply repeating predigested words and sentences". This is congruent with the results of other studies based on the perception of both students and teachers that yield positive reactions towards ASR (CHEN, 2011; INCEOGLU; LIM; CHEN, 2020). In addition, this positive reaction might be connected to the fact that learners comprehend the importance of pronunciation for successful communication, being keen to use technological resources to improve it (ROGERSON-REVELL, 2021).

All in all, the potential of ASR for pronunciation improvement in second language acquisition is enormous. Although "ASR will never be 100% accurate" (KNILL et al., 2018, p. 1641), teachers and learners can focus on the strengths of this technology to fulfill specific learning goals. ASR can provide learners with the opportunity to produce oral output and therefore practice pronunciation and speaking skills. Chapelle and Jamieson (2008, p. 151) define the latter skill as "a fast-paced mental and physical activity that requires the speaker to process linguistic knowledge automatically". Therefore, ASR could also be used for practicing such skill.

3.2 Text-to-speech and its affordances for L2 pronunciation improvement

Considering the central role of oral input for the development of perceptual skills and infrequent exposure to it in instructional contexts, as discussed in section 2, pronunciation teaching also needs to focus on the level of perception. However, according to Darcy (2018), to avoid class time overload

and monotonous tasks, three aspects must be considered in perceptual listening activities: 1) contextualized and repeated links to the vocabulary that is already being studied (rather than in unknown words such as minimal pairs); 2) variability in the input, by exposing learners to different accents, voices, and speech rates, for instance; and 3) multimodality, that is, presenting the language instances in more than one modality (e.g., oral input plus written input).

Following this perspective, TTS, also called speech synthesis, is one specific technology to enhance perceptual knowledge. Straightforwardly, Handley (2013, p. 5846) defines speech synthesis as "the process of making the computer talk", and it is a widely used technology that automatically generates synthesized speech from units of written text displayed on a screen (LIAKIN; CARDOSO; LIAKINA, 2017). Currently, there is a variety of TTS programs either as free or paid versions, which may differ in some of their functions, as more sophisticated versions usually offer voice options (e.g., male or female, native or nonnative-like, different language varieties and accents), more interaction to the user, such as highlighting each word being read aloud, and more access to other types of files (e.g., PDFs, e-books, and web documents) (MOON, 2012).

Although this technology is not recent, it has received significant improvement over the last few years in order to become more similar to natural human speech (HANDLEY, 2013; LIAKIN; CARDOSO; LIAKINA, 2017; MOON, 2012). Along with the advances, TTS has attracted scholars' attention as a potential pedagogical tool, especially concerning whether and how it could assist L2 classes due to potential uses for different aspects of language learning. For example, Moon (2012) proposes that TTS provides several opportunities for practicing all four skills: writing, reading, listening, and speaking. For illustration, the author suggests its use for learners to 1) revise their written texts; 2) create and download audio versions from any text for listening to a topic of interest; 3) adjust the speed and pace of the audio in order to facilitate understanding of the content being read; 4) elaborate and practice dialogues with different English accents; and 5) check on the pronunciation of individual words.

In addition, Handley (2013) points out that TTS has more applicable and apparent benefits for some language domains. As an example, the technology can help in exercises to reinforce the relationship between graphemes and phonemes of the target language, which could supplement writing and reading abilities, and, therefore, widen learners' vocabulary knowledge.

More recently, some attention has been given to analyzing TTS speech quality in research (LIAKIN; CARDOSO; LIAKINA, 2017). For instance, Cardoso, Smith, and Garcia Fuentes (2015) investigated if it could generate speech like human performance under four aspects: comprehensibility, naturalness, pronunciation accuracy, and intelligibility. To answer the question, the authors recruited 15 undergraduate students to rate oral samples provided either by TTS or by human recordings on two conditions: only sentences or within a story. Participants were also engaged in an identification task to focus on the past form of regular verbs in English, in which they should judge if the sentences contained or not the target grammar feature. A sequence of paired t-tests sample was run, revealing a significant difference in the overall scores of human and TTS oral productions on both conditions (story and sentences). However, the tests showed no significant differences between the samples in the identification task. Likewise, it is worthy of mention that, despite not achieving an equal baseline as the human recordings, participants assigned relatively high scores to the TTS samples under the dimensions of comprehensibility, pronunciation accuracy, and intelligibility in both conditions. These results indicate, therefore, that the synthesized voice generated by the technology can currently constitute an adequate source of spoken input for L2 learners.

As cited by Liakin, Cardoso, and Liakina (2017), the authors' findings indicate that the quality of the synthesized voice can resemble the human voice, allowing its use as a pronunciation model. Grimshaw, Bione Alves, and Cardoso (2018) obtained similar results when analyzing the output of five different TTS applications. The study showed that the participants' overall ratings for comprehensibility were relatively high, although scores for the naturalness dimension were well below comprehensibility. In addition, two aspects can be pointed out from this study. First, as language users become more familiar with the specific synthesized voice, they can perceive speech as more understandable and natural (BIONE ALVES, 2017 apud GRIMSHAW; BIONE ALVES; CARDOSO,

2018). This implies that, if implemented on a recurring basis as a pedagogical resource, the output generated by TTS may be perceived with better quality. Second, considering the variability still present among TTS applications, it might be necessary to assess the speech quality of whichever application before using it as a pedagogical tool.

The study conducted by Liakin, Cardoso, and Liakina (2017), which aimed to investigate students' perception of both ASR and TTS as pedagogical tools, equally supports the use of TTS as an additional source of oral L2 input. The authors reported a previous experiment in which participants were divided into different groups and engaged in weekly teacher-led or technology-led practices (with ASR or TTS). After the experimental period, the authors asked the participants from the ASR group (n=14) and TTS group (n=9) to take part in a survey and an interview, inquiring about their attitudes and perceptions of the tools for pronunciation practice. The data from the survey was quantitatively analyzed through descriptive statistics and revealed, overall, that "the participants evaluated positively their experience with these two mobile speech technologies and, more importantly, they found them useful, practical and helpful for their own learning" (LIAKIN; CARDOSO; LIAKINA, 2017, p. 21). Likewise, the qualitative analysis from the interviews corroborates most findings from the quantitative data, evidencing the learners' positive attitudes towards the use of the tools once again.

With more specific reference to TTS technology, participants' responses collected in the interview acknowledged some of the aforementioned possibilities to take advantage of the tool, such as extensive listening and oral comprehension practices. Besides the perception practice, some learners also noticed an improvement in their pronunciation after using TTS. According to the authors, such a gain is "explained by the fact that the app increased their exposure and access to the correct pronunciation model" (LIAKIN; CARDOSO; LIAKINA, 2017, p. 24).

These research findings indicate that TTS technology seems ready to be used in L2 classes. It has much to offer for pronunciation practice, "particularly as a supplemental source of input which can cater for learners' individual needs and interests" (CARDOSO, 2018, p. 112). Moreover, Cardoso (2018) has empirically attested to such a claim in a study designed to analyze the use of TTS in learning the pronunciations of the past tense marker of regular verbs in English. The participants of the study received a four-week treatment similar in numbers of activities but different in the feedback provided, either by TTS or by the language teacher. In a pre-post-test design, the author observed that both groups (TTS and Non-TTS) obtained similar scores, showing that TTS could be implemented as an out-of-class teaching tool to "increase in-class time so that teachers and students could focus on other important tasks" (CARDOSO; SMITH; GARCIA FUENTES, 2015, p. 21). The author, however, stresses that in this experiment, the use of the TTS enhanced pronunciation at the level of perception. A possible positive effect of this enhanced pronunciation is that learners can engage in other speaking activities so as to transfer the obtained perceptual knowledge into production.

In this regard, Eksi and Yesilcinar (2016) were able to report production gains after engaging participants in a rehearsing session to practice an oral presentation with the help of TTS programs. The authors recruited 43 EFL teacher trainees to participate in an experimental instruction, preceded and followed by a testing section. At the end of the experiment, the participants also filled in a post reflection questionnaire to assess their impressions regarding using the tools in their self-studies. The pre- and post-tests consisted of five-minute-long oral presentations delivered by each teacher trainee, which were recorded and rated according to five aspects (fluency, pronunciation and accent, vocabulary, accuracy, and content). The scores obtained from the tests reveal a significant difference when analyzing the teacher trainees' scores and their overall pronunciation and fluency performance in the post-test, suggesting that TTS was indeed helpful to promote pronunciation improvements. Likewise, the participants' reflection on the questionnaire acknowledged the user-friendly interface of the application and their willingness to make use of this technology, as most of them expressed the intention to keep using it in self-studies.

Concerning attitude towards using TTS, students seem to be willing to use it in their learning process. BIONE ALVES, Grimshaw, and Cardoso (2016) analyzed the perceptions of fifteen Brazilian learners of EFL about the quality of texts generated by TTS and concluded that "EFL learners have overall positive attitudes towards the pedagogical use of TTS, and that they would like to use the

technology as a learning tool" (BIONE ALVES; GRIMSHAW; CARDOSO, 2016, p. 50).

Given these research findings, language teachers and instructors have a lot to benefit from TTS. As seen, this technology can provide limitless oral input in the target language (CARDOSO; SMITH; GARCIA FUENTES, 2015), promote efficient and personalized feedback (CARDOSO, 2018), raise learners' awareness to specific features and forms of the L2 (LIAKIN; CARDOSO; LIAKINA, 2017; GOMES; CARDOSO; LUCENA, 2018; CARDOSO, 2018) and increase learners' autonomy in their own phonological development (MOON, 2012; LIAKIN; CARDOSO; LIAKINA, 2017).

However, in order to enhance gains at the production level as well, it would be more appropriate to combine the use of TTS with other tools that elicit speech production on the users' part. Hence, Liakin, Cardoso, and Liakina (2017) suggest combining both ASR and TTS as an "anytime anywhere mobile learning setting" seems to be a promising proposal to facilitate pronunciation improvement. In the following section, we further discuss the combination of these technologies and suggest some digital resources to explore these technologies for pedagogical purposes.

3.3 Integrating ASR and TTS to pronunciation teaching

The previous sections demonstrated the affordances of both ASR and TTS technologies for pronunciation teaching and learning. All things considered, it is possible to reason that these technologies can foster pronunciation improvement inside (under the teacher's guidance) and outside (autonomously by the learner) the classroom. Liakin, Cardoso, and Liakina (2017) argue that researchers have only started to explore the pedagogical uses of TTS and ASR in L2 education; notwithstanding, the available studies so far suggest positive results as a classroom instruction complement after their extended use. Similarly, Levis and Suvorov (2013b) indicate that the connections between ASR and text-to-speech software have not been fully explored, and that they can have promising results for non-native speech applications.

Considering the learner's side, Golonka et al. (2014) state that technological innovations can provide learners with interaction opportunities, feedback, and more contact with the target language, besides increasing their motivation and interest. Such a statement is also congruent with students' perceptions regarding ASR and TTS programs. In this regard, Liakin, Cardoso, and Liakina (2017) observed comparative results when investigating learners' impressions on the use of both technologies. According to the authors, the participants not only recognized the pedagogical importance of ASR and TTS but also enjoyed the mobile-enhanced learning environment afforded by them.

Turning our focus to pronunciation teaching, these two technologies can be combined to provide learners with relevant oral input (TTS), opportunities to practice oral output (ASR), and the opportunity to check their production in relation to the perceptual knowledge they have acquired. Thus, this blended usage can be a powerful resource for L2 language teachers. For instance, it is possible to use these digital resources to offer relevant extra-class activities once they can make it easier to assign beneficial self-administered pronunciation tasks (MUNRO; DERWING, 2015).

Another relevant pedagogical implication is that ASR and TTS can facilitate integrating the pronunciation component in all sorts of language courses, as defended by Darcy, Rocca, and Hancock (2021) and Gordon and Darcy (2016). In line with the authors' view, pronunciation gains can be achieved even in short periods of instruction, without the (actual) need to devote a whole course to pronunciation instruction. Rather, it is viable to allocate some focus on both segmental and suprasegmental features within their communicative classes (DARCY; ROCCA; HANCOCK, 2021). As a result, having recurrent pronunciation practice is likely to yield significant gains on the learners' pronunciation skills. Furthermore, ASR and TTS programs can be employed whenever specific sounds and features are taught in class to back up the practice of the target aspects. In this way, their use can be adapted for different learning units throughout a course or school year.

Moreover, both programs allow for individualized instruction, a core standpoint of the Intelligibility Principle for pronunciation teaching (THOMSON; DERWING, 2014; MUNRO; DERWING, 2015). In agreement with the principle, it is advisable to evaluate the instruction needs according to the learner's specific demands, especially concerning the aspects likely to hinder their intelligibility. Hence, in a classroom environment, resorting to digital resources is an alternative way to tailor a more

individualized instruction. This is especially relevant with regard to the time constraints teachers face in complying with the regular curriculum, enabling them to devote more class time to the difficulties shared by a greater number of students (ROCCAMO, 2014; MUNRO; DERWING, 2015).

In a more practical vein, there are many digital resources available for free that use both ASR and TTS technology. A simple way to use them is by voice web search (e.g., https://www.google.com). The intended message is dictated on the textbox after clicking on the microphone icon, and the result is read out-loud by the TTS tool. Another possibility is an online translator (e.g., https://translate.google.com/). Instead of typing words, it is possible to dictate them to the program, or discover the translation into the target language and listen to its pronunciation (for a more detailed guideline, see (CARRIER, 2017). For instruction focused on form[2], an immersive reader (e.g., Microsoft Word[3], Microsoft Edge[4]) can provide the target form while an ASR-based dictation tool (e.g., https://dictation.io/, https://speechnotes.com) can be used for the output drills[5]. Thus, teachers can invest more time in encouraging communicative activities with the low-level basic pronunciation drills done with the ASR software (KIM, 2006).

Similarly, VoiceNotebook's[6] pronunciation practice page not only integrates both ASR and TTS in a single webpage but also includes a playback feature. It also displays a comparison between a target text passage with the transcribed utterances from the ASR tool, proving better feedback for the learner to focus on their main difficulties.

It is worth mentioning that the ASR and TTS as well as their pedagogical implications can be applied for different learning contexts and different languages. As Henrichsen (2021) indicates, ASR technology can convert speech into text in over one hundred languages. Moreover, the aforesaid ways to use the tools can be applied to different target languages if supported by the applications.

As a final statement, based on the definition of efficient pronunciation teaching by Darcy (2018), as aforementioned, it is possible to fulfill all those 'ingredients' by combining both technologies: ASR can be an integrative part of different explicit and communicative activities providing learners with endless opportunities of producing oral output, TTS can be used to develop learners' perception and ASR can provide automatic explicit feedback at the learner's pace. Hence, the combined use of both technologies can be an attractive option for the long sought pronunciation improvement.

4 Final remarks

The discussion in this paper aimed to present the affordances of ASR and TTS for L2 pronunciation improvement. For this purpose, a review of literature on previous empirical studies was carried out, with research investigating learners and teachers' perceptions and the use of ASR and TTS as a pedagogical tool for pronunciation practice.

Given that pronunciation inaccuracies can result in communication breakdowns, language users may have to improve their pronunciation abilities in order to deliver and comprehend L2 speech more efficiently. However, as seen, it is not so simple to tailor pronunciation instruction in a classroom environment, mainly due to time constraints. Hence, the technological resources presented – ARS and TTS – can facilitate pronunciation practice, in terms of both perception and production. Unfortunately, even the most advanced applications for pronunciation practice "are still lacking explicit feedback for acquisition and assessment of foreign language suprasegmentals" (BOGACH et al., 2021, p. 3). Therefore, these tools might be more appropriate for practicing segmental features of the language if used as an autonomous learning tool without the guidance of a teacher.

As the aforementioned studies suggest, technology can result in considerable benefits to learners, mainly as a supplement to pronunciation teaching. Nonetheless, these benefits can only be achieved if the teacher is aware of "what their students need, and if they use tools that have been shown

2 Focus on form refers to drawing learners' attention to linguistic elements encountered in lesson (LONG, 1991).
3 https://support.microsoft.com/en-us/office/listen-to-your-word-documents-5a2de7f3-1ef4-4795-b24e-64fc2731b001.Accessed:Dec12th,2021.
4 https://support.microsoft.com/en-us/topic/use-immersive-reader-in-microsoft-edge-78a7a17d-52e1-47ee-b0ac-eff8539015e1. Accessed: Dec 12th, 2021.
5 Asking students to repeat individually an intended utterance to practice specific linguistic elements (HARMER, 2012).
6 https://voicenotebook.com/prononce.php

to be effective" (DARCY, 2018, p. 326) Thus, the use of technological resources is not to replace the important role of the teacher. Rather, it is a way of giving learners the opportunity to focus on their specific difficulties and receive personalized feedback while becoming more autonomous in their learning process, meaning that "learning is not limited to the classroom context" (CARLET; KIVISTÖ-DE SOUZA, 2018, p. 104).

All being said, these technologies hold a bright future in CALL. However, "educational technology is only as good as the humans behind it" (ROGERSON-REVELL, 2021, p. 201). Hence, this paper sought to shed some light on the usage of two prodigious technologies, demonstrating their effectiveness, qualities, and pitfalls to assist teachers in their pedagogical decisions to meet their students' needs. Notwithstanding, the best tools may "not necessarily [be] those that seem newest, coolest, or flashiest" (YOSHIDA, 2018, p. 209) but the most adequate for the learning purpose. Moreover, teachers will be demanded more and more to make use of technological resources, implying that they may also have to adapt their practices to deal with these demands (MENEZES, 2019). A clear picture of the current possibilities available is, thus, a relevant takeaway.

References

ASHWELL, T.; ELAM, J. R. How accurately can the Google Web Speech API recognize and transcribe Japanese L2 English learners' oral production? *The JALT CALL Journal*, v. 13, n. 1, p. 59–76, Apr. 2017. DOI: 10.29140/jaltcall.v13n1.212. Available from: <https://www.castledown.com/journals/jaltcall/article/?reference=j212>. Visited on: 9 Feb. 2022.

BIONE ALVES, T. *Synthetic voices in the foreign language context*. 2017. Master's Thesis – Concordia University, Montreal, CA.

BIONE ALVES, T.; GRIMSHAW, J.; CARDOSO, W. An evaluation of text-to-speech synthesizers in the foreign language classroom: learners' perceptions. In: PAPADIMA-SOPHOCLEOUS, Salomi; BRADLEY, Linda; THOUËSNY, Sylvie (Eds.). *CALL communities and culture – short papers from EUROCALL 2016*. [S.l.]: Research-publishing.net, Dec. 2016. p. 50–54. DOI: 10.14705/rpnet.2016.eurocall2016.537. Available from: <https://research-publishing.net/manuscript?10.14705/rpnet.2016.eurocall2016.537>. Visited on: 9 Feb. 2022.

BOGACH, N. et al. Speech Processing for Language Learning: A Practical Approach to Computer-Assisted Pronunciation Teaching. *Electronics*, v. 10, n. 3, p. 235, Jan. 2021. DOI: 10.3390/electronics10030235. Available from: <https://www.mdpi.com/2079-9292/10/3/235>. Visited on: 9 Feb. 2022.

CARDOSO, W. Learning L2 pronunciation with a text-to-speech synthesizer. In: FUTURE-PROOF CALL: language learning as exploration and encounters – short papers from EUROCALL 2018. [S.l.]: Research-publishing.net, Dec. 2018. p. 16–21. DOI: 10.14705/rpnet.2018.26.806. Available from: <https://research-publishing.net/manuscript?10.14705/rpnet.2018.26.806>. Visited on: 9 Feb. 2022.

CARDOSO, W.; SMITH, G.; GARCIA FUENTES, C. Evaluating text-to-speech synthesizers. In: CRITICAL CALL – Proceedings of the 2015 EUROCALL Conference, Padova, Italy. [S.l.]: Research-publishing.net, Dec. 2015. p. 108–113. DOI: 10.14705/rpnet.2015.000318. Available from: <https://research-publishing.net/manuscript?10.14705/rpnet.2015.000318>. Visited on: 9 Feb. 2022.

CARLET, A.; KIVISTÖ-DE SOUZA, H. Improving L2 pronunciation inside and outside the classroom. *Ilha do Desterro A Journal of English Language, Literatures in English and Cultural Studies*, v. 71, n. 3, p. 99–124, Sept. 2018. DOI: 10.5007/2175-8026.2018v71n3p99. Available from: <https://periodicos.ufsc.br/index.php/desterro/article/view/2175-8026.2018v71n3p99>. Visited on: 9 Feb. 2022.

CARRIER, M. Automated Speech Recognition in language learning: Potential models, benefits and impact. *Training Language and Culture*, v. 1, n. 1, p. 46–61, Feb. 2017. DOI: 10.29366/2017tlc.1.1.3. Available from: <http://rudn.tlcjournal.org/issues/1(1)-03.html>. Visited on: 9 Feb. 2022.

CHAPELLE, C.; JAMIESON, J. *Tips for teaching with CALL: practical approaches to computer-assisted language learning*. White Plains, NY: Pearson Education, 2008. (Tips on teaching).

CHEN, H. H.-J. Developing and evaluating an oral skills training website supported by automatic speech recognition technology. en. *ReCALL*, v. 23, n. 1, p. 59–78, Jan. 2011. DOI: 10.1017/S0958344010000285. Available from: <https://www.cambridge.org/core/product/identifier/S0958344010000285/type/journal_article>. Visited on: 9 Feb. 2022.

CHUN, D.; KERN, R.; SMITH, B. Technology in Language Use, Language Teaching, and Language Learning. *The Modern Language Journal*, v. 100, S1, p. 64–80, Jan. 2016. DOI: 10.1111/modl.12302. Available from: <https://onlinelibrary.wiley.com/doi/10.1111/modl.12302>. Visited on: 9 Feb. 2022.

DARCY, I. Powerful and Effective Pronunciation Instruction: How Can We Achieve It? en. *CATESOL Journal*, v. 30, n. 1, p. 13–45, 2018. Available from: <https://eric.ed.gov/?id=EJ1174218>. Visited on: 9 Feb. 2022.

DARCY, I.; ROCCA, B.; HANCOCK, Z. A Window into the Classroom: How Teachers Integrate Pronunciation Instruction. *RELC Journal*, v. 52, n. 1, p. 110–127, Apr. 2021. DOI: 10.1177/0033688220964269. Available from: <http://journals.sagepub.com/doi/10.1177/0033688220964269>. Visited on: 9 Feb. 2022.

DAVIES, G. Computer-Assisted Language Education. In: BERNS, M.; BROWN, C. (Eds.). *Concise Encyclopedia of Applied Linguistics*. Oxford: Elsevier, 2006. p. 261–271.

DEKEYSER, R. Skill Acquisition Theory. In: VANPATTEN, B.; WILLIAMS, J. (Eds.). *Theories in Second Language Acquisition: An Introduction*. New York and London: Routhledge, 2015. p. 97–113.

DEMENKO, G.; WAGNER, A.; CYLWIK, N. The Use of Speech Technology in Foreign Language Pronunciation Training. *Archives of Acoustics*, v. 35, n. 3, p. 309–329, Sept. 2010. DOI: 10.2478/v10168-010-0027-z. Available from: <https://content.sciendo.com/doi/10.2478/v10168-010-0027-z>. Visited on: 9 Feb. 2022.

DERWING, T. M. The efficacy of pronunciation instruction. In: KANG, O.; THOMSON, R. I.; MURPHY, J. M. (Eds.). *he Routledge Handbook of Contemporary English Pronunciation*. Milton Park: Routledge, 2018. p. 320–334.

DERWING, T. M.; MUNRO, Murray J.; CARBONARO, M. Does Popular Speech Recognition Software Work with ESL Speech? *TESOL Quarterly*, v. 34, n. 3, p. 592, 2000. DOI: 10.2307/3587748. Available from: <https://www.jstor.org/stable/3587748?origin=crossref>. Visited on: 9 Feb. 2022.

DIZON, G. Evaluating Intelligent Personal Assistants for L2 Listening and Speaking Development. *Language Learning & Technology*, v. 24, p. 16–26, 2020.

DIZON, G.; TANG, D. Intelligent personal assistants for autonomous second language learning: An investigation of Alexa. *The JALT CALL Journal*, v. 16, n. 2, p. 107–120, Aug. 2020. DOI: 10.29140/jaltcall.v16n2.273. Available from: <https://www.castledown.com/journals/jaltcall/article/?reference=273>. Visited on: 9 Feb. 2022.

EKSI, G. Y.; YESILCINAR, S. An Investigation of the Effectiveness of Online Text-to-Speech Tools in Improving EFL Teacher Trainees' Pronunciation. *English Language Teaching*, v. 9, n. 2, p. 205, Jan. 2016. DOI: 10.5539/elt.v9n2p205. Available from: <http://www.ccsenet.org/journal/index.php/elt/article/view/56606>. Visited on: 9 Feb. 2022.

GOLONKA, E. M. et al. Technologies for foreign language learning: a review of technology types and their effectiveness. *Computer Assisted Language Learning*, v. 27, n. 1, p. 70–105, Feb. 2014. DOI: 10.1080/09588221.2012.700315. Available from: <http://www.tandfonline.com/doi/abs/10.1080/09588221.2012.700315>. Visited on: 9 Feb. 2022.

GOMES, A. A. de A.; CARDOSO, W.; LUCENA, R. M. de. Can TTS help L2 learners develop their phonological awareness? In: FUTURE-PROOF Call: language learning as exploration and encounters – short papers from EUROCALL 2018. [S.l.: s.n.], 2018. p. 29–34. DOI: http://dx.doi.org/10.14705/rpnet.2018.26.808. Available from: <https://research-publishing.net>. Visited on: 9 Feb. 2022.

GORDON, J.; DARCY, I. The development of comprehensible speech in L2 learners: A classroom study on the effects of short-term pronunciation instruction. *Journal of Second Language Pronunciation*, v. 2, n. 1, p. 56–92, Mar. 2016. DOI: 10.1075/jslp.2.1.03gor. Available from: <http://www.jbe-platform.com/content/journals/10.1075/jslp.2.1.03gor>. Visited on: 9 Feb. 2022.

GRASS, S. M.; MACKEY, A. Input, Interaction, and Output in Second Language Acquisition. In: VANPATTEN, B.; WILLIAMS, J. (Eds.). *Theories in second language acquisition: an introduction*. Second Edition. New York: Routledge, 2015. (Second Language Acquisition Research Series).

GRIMSHAW, J.; BIONE ALVES, T.; CARDOSO, W. Who's got talent? Comparing TTS systems for comprehensibility, naturalness, and intelligibility. In: FUTURE-PROOF CALL: language learning as exploration and encounters – short papers from EUROCALL 2018. [S.l.]: Research-publishing.net, Dec. 2018. p. 83–88. DOI: 10.14705/rpnet.2018.26.817. Available from: <https://research-publishing.net/manuscript?10.14705/rpnet.2018.26.817>. Visited on: 9 Feb. 2022.

HANDLEY, Z. Text-to-Speech Synthesis in Computer-Assisted Language Learning. In: CHAPELLE, C. A. (Ed.). *The encyclopedia of applied linguistics*. New York: Wiley-Blackwell, 2013. p. 5846–5851.

HARMER, J. *Essential teacher knowledge. Buch*. Harlow: Pearson Education, 2012. (Always learning).

HENRICHSEN, L. E. An Illustrated Taxonomy of Online CAPT Resources. *RELC Journal*, v. 52, n. 1, p. 179–188, Apr. 2021. DOI: 10.1177/0033688220954560. Available from: <http://journals.sagepub.com/doi/10.1177/0033688220954560>. Visited on: 9 Feb. 2022.

INCEOGLU, S.; LIM, H.; CHEN, W.-H. ASR for EFL Pronunciation Practice: Segmental Development and Learners' Beliefs. *The Journal of AsiaTEFL*, v. 17, n. 3, p. 824–840, Sept. 2020. DOI: 10.18823/asiatefl.2020.17.3.5.824. Available from: <http://journal.asiatefl.org/main/main.php?inx_journals=64&inx_contents=842&submode=3&PageMode=JournalView&s_title=ASR_for_EFL_Pronunciation_Practice_Segmental_Development_and_Learners_Beliefs>. Visited on: 9 Feb. 2022.

JURAFSKY, D.; MARTIN, J. H. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Uttar Pradesh (India): Pearson, 2000.

KIM, I.-S. Automatic Speech Recognition: Reliability and Pedagogical Implications for Teaching Pronunciation. *Journal of Educational Technology & Society*, v. 9, n. 1, p. 322–334, 2006. Available from: <https://www.jstor.org/stable/jeductechsoci.9.1.322>. Visited on: 9 Feb. 2022.

KNILL, K. et al. Impact of ASR Performance on Free Speaking Language Assessment. In: INTERSPEECH 2018. [S.l.]: ISCA, Sept. 2018. p. 1641–1645. DOI: 10.21437/Interspeech.2018-1312. Available from: <https://www.isca-speech.org/archive/interspeech_2018/knill18_interspeech.html>. Visited on: 9 Feb. 2022.

LEE, B.; PLONSKY, L.; SAITO, K. The effects of perception- vs. production-based pronunciation instruction. *System*, v. 88, p. 102185, Feb. 2020. DOI: 10.1016/j.system.2019.102185. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0346251X19305196>. Visited on: 9 Feb. 2022.

LEVIS, J.; SUVOROV, R. Automatic speech recognition. In: CHAPELLE, C. (Ed.). *The encyclopedia of applied linguistics*. [S.l.: s.n.], 2013. Available from: <http://www.credoreference.com/book/wileyenapl>. Visited on: 9 Feb. 2022.

LEVIS, J.; SUVOROV, R. Automatic speech recognition. In: CHAPELLE, C. (Ed.). *The encyclopedia of applied linguistics*. [S.l.: s.n.], 2013. Available from: <http://www.credoreference.com/book/wileyenapl>. Visited on: 9 Feb. 2022.

LIAKIN, D.; CARDOSO, W.; LIAKINA, N. Mobilizing Instruction in a Second-Language Context: Learners' Perceptions of Two Speech Technologies. *Languages*, v. 2, n. 3, p. 11, July 2017. DOI: 10.3390/languages2030011. Available from: <http://www.mdpi.com/2226-471X/2/3/11>. Visited on: 9 Feb. 2022.

LONG, M. H. Focus on form: A design feature in language teaching methodology. In: DE BOT, K.; GINSBERG, R.; KRAMSCH, C. (Eds.). *Foreign language research in cross-cultural perspective*. Amsterdam: John Benjamins, 1991. p. 39. Available from: <https://benjamins.com/catalog/sibil.2.07lon>. Visited on: 9 Feb. 2022.

MARTINS, C. B.; MOREIRA, H. O campo CALL (Computer Assisted Language Learning): definições, escopo e abrangência. *Calidoscópio*, v. 10, n. 3, p. 247–255, Dec. 2012. DOI: 10.4013/cld.2012.103.01. Available from: <http://revistas.unisinos.br/index.php/calidoscopio/article/view/3254>. Visited on: 9 Feb. 2022.

MCCROCKLIN, S.; EDALATISHAMS, I. Revisiting Popular Speech Recognition Software for ESL Speech. *TESOL Quarterly*, v. 54, n. 4, p. 1086–1097, Dec. 2020. DOI: 10.1002/tesq.3006. Available from: <https://onlinelibrary.wiley.com/doi/10.1002/tesq.3006>. Visited on: 9 Feb. 2022.

MENEZES, V. Tecnologias digitais no ensino de línguas: passado, presente e futuro. *Revista da ABRALIN*, Aug. 2019. DOI: 10.25189/rabralin.v18i1.1323. Available from: <https://revista.abralin.org/index.php/abralin/article/view/1323>. Visited on: 9 Feb. 2022.

MOON, D. Web-Based Text-to-Speech Technologies in Foreign Language Learning: Opportunities and Challenges. In: KIM, T.-H. et al. (Eds.). *Computer Applications for Database, Education, and Ubiquitous Computing*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. v. 352. p. 120–125. DOI: 10.1007/978-3-642-35603-2_19. Available from: <http://link.springer.com/10.1007/978-3-642-35603-2_19>. Visited on: 9 Feb. 2022.

MOUSSALLI, S.; CARDOSO, W. Intelligent personal assistants: can they understand and be understood by accented L2 learners? *Computer Assisted Language Learning*, v. 33, n. 8, p. 865–890, Nov. 2020. DOI: 10.1080/09588221.2019.1595664. Available from: <https://www.tandfonline.com/doi/full/10.1080/09588221.2019.1595664>. Visited on: 9 Feb. 2022.

MROZ, A. Seeing how people hear you: French learners experiencing intelligibility through automatic speech recognition. *Foreign Language Annals*, v. 51, n. 3, p. 617–637, Sept. 2018. DOI: 10.1111/flan.12348. Available from: <https://onlinelibrary.wiley.com/doi/10.1111/flan.12348>. Visited on: 9 Feb. 2022.

MUNOZ, C. Symmetries and Asymmetries of Age Effects in Naturalistic and Instructed L2 Learning. *Applied Linguistics*, v. 29, n. 4, p. 578–596, Jan. 2008. DOI: 10.1093/applin/amm056. Available from: <https://academic.oup.com/applij/article-lookup/doi/10.1093/applin/amm056>. Visited on: 9 Feb. 2022.

MUNRO, M. J.; DERWING, T. M. Foreign Accent, Comprehensibility, and Intelligibility in the Speech of Second Language Learners. *Language Learning*, v. 45, n. 1, p. 73–97, Mar. 1995. DOI: 10.1111/j.1467-1770.1995.tb00963.x. Available from: <https://onlinelibrary.wiley.com/doi/10.1111/j.1467-1770.1995.tb00963.x>. Visited on: 9 Feb. 2022.

MUNRO, M. J.; DERWING, T. M. Intelligibility in Research and Practice: Teaching Priorities. In: REED, M.; LEVIS, J. M. (Eds.). *The Handbook of English Pronunciation*. 1. ed. [S.l.]: Wiley, May 2015. p. 375–396. DOI: 10.1002/9781118346952.ch21. Available from: <https://onlinelibrary.wiley.com/doi/10.1002/9781118346952.ch21>. Visited on: 9 Feb. 2022.

MUNRO, M. J.; DERWING, T. M.; MORTON, S. L. THE MUTUAL INTELLIGIBILITY OF L2 SPEECH. *Studies in Second Language Acquisition*, v. 28, n. 01, Mar. 2006. DOI: 10.1017/S0272263106060049. Available from: <http://www.journals.cambridge.org/abstract_S0272263106060049>. Visited on: 9 Feb. 2022.

O'BRIEN, M. G. et al. Directions for the future of technology in pronunciation research and teaching. *Journal of Second Language Pronunciation*, v. 4, n. 2, p. 182–207, Dec. 2018. DOI: 10.1075/jslp.17001.obr. Available from: <http://www.jbe-platform.com/content/journals/10.1075/jslp.17001.obr>. Visited on: 9 Feb. 2022.

ORTEGA, L. *Understanding second language acquisition*. London: Routledge, 2009. (Understanding language series).

PENNINGTON, M. C.; ROGERSON-REVELL, P. *English Pronunciation Teaching and Research: Contemporary Perspectives*. London: Palgrave Macmillan UK, 2019. DOI: 10.1057/978-1-137-47677-7. Available from: <http://link.springer.com/10.1057/978-1-137-47677-7>. Visited on: 9 Feb. 2022.

ROCCAMO, A. Effective pronunciation instruction in basic language classrooms: A modular approach. In: LEVIS, J.; MCCROCKLIN, S. (Eds.). *Proceedings of the 5th Pronunciation in Second Language Learning and Teaching Conference*. Ames, IA: Iowa State University, 2014. p. 183–189.

ROGERSON-REVELL, P. M. Computer-Assisted Pronunciation Training (CAPT): Current Issues and Future Directions. en. *RELC Journal*, v. 52, n. 1, p. 189–205, Apr. 2021. DOI: 10.1177/0033688220977406. Available from: <http://journals.sagepub.com/doi/10.1177/0033688220977406>. Visited on: 9 Feb. 2022.

SICOLA, L.; DARCY, I. Integrating Pronunciation into the Language Classroom. In: REED, M.; LEVIS, J. M. (Eds.). *The Handbook of English Pronunciation*. 1. ed. [S.l.]: Wiley, May 2015. p. 471–487. DOI: 10.1002/9781118346952.ch26. Available from: <https://onlinelibrary.wiley.com/doi/10.1002/9781118346952.ch26>. Visited on: 9 Feb. 2022.

SLABAKOVA, R. *Second Language Acquisition*. Oxford: Oxford University Press, 2016.

THOMSON, R. I.; DERWING, T. M. The effectiveness of L2 pronunciation instruction: a narrative review. *Applied Linguistics*, v. 36, n. 3, p. 326–344, July 2014. DOI: 10.1093/applin/amu076. Available from: <https://academic.oup.com/applij/article-lookup/doi/10.1093/applin/amu076>. Visited on: 9 Feb. 2022.

TYLER, M. PAM-L2 and phonological category acquisition in the foreign language classroom. In: [s.l.: s.n.], May 2019. p. 607–630.

VANPATTEN, B. Processing matters in input enhancement. In: PISKE, T.; YOUNG-SCHOLTEN, M. (Eds.). *Input Matters in SLA*. [S.l.]: Multilingual Matters, Dec. 2008. p. 47–61. DOI: 10.21832/9781847691118-005. Available from: <https://www.degruyter.com/document/doi/10.21832/9781847691118-005/html>. Visited on: 9 Feb. 2022.

VANPATTEN, B.; SMITH, M.; BENATI, A. G. *Key questions in second language acquisition: an introduction*. New York: Cambridge University Press, 2019.

YAVAŞ, M. *Applied English phonology*. 2nd ed. Oxford ; Malden, MA: Wiley-Blackwell, 2011.

YOSHIDA, M. T. Choosing Technology Tools to Meet Pronunciation Teaching and Learning Goals. *CATESOL Journal*, v. 30, n. 1, p. 195–212, 2018. Available from: <https://eric.ed.gov/?id=EJ1174226>. Visited on: 9 Feb. 2022.

ZHANG, R.; YUAN, Z.-M. Examining the effects of explicit pronunciation instruction on the development of l2 pronunciation. *Studies in Second Language Acquisition*, v. 42, n. 4, p. 905–918, Sept. 2020. DOI: 10.1017/S0272263120000121. Available from: <https://www.cambridge.org/core/product/identifier/S0272263120000121/type/journal_article>. Visited on: 9 Feb. 2022.

Author contributions

William Gottardi: Writing – original draft, Writing – review and editing; **Janaina Fernanda de Almeida**: Writing – original draft, Writing – review and editing; **Celso Henrique Soufen Tumolo**: Writing – original draft, Writing – review and editing .