Synthetic voices in the foreign language context

Tiago Bione Alves

A Thesis

in

The Department

of

Education

Presented in Partial Fulfillment of the Requirements

for the Degree of Master of Arts (Applied Linguistics) at

Concordia University

Montreal, Quebec, Canada

September 2017

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: Tiago Bione Alves

Entitled: Synthetic voices in the foreign language context

and submitted in partial fulfillment of the requirements for the degree of

Master of Arts (Applied Linguistics)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final Examining Committee:

_____ Chair

Guilherme Garcia

_____ Examiner

Laura Collins

_____ Examiner

Denis Liakin

_____ Supervisor

Walcir Cardoso

Approved by _____

Chair of Department or Graduate Program Director

_____2017 _____

Dean of Faculty

ABSTRACT

Synthetic voices in the foreign language context

Tiago Bione Alves

Second language (L2) researchers and practitioners have explored the pedagogical capabilities of text-to-speech synthesizers (TTS) for their potential to enhance the acquisition of writing (Kirstein, 2006), vocabulary and reading (Proctor, Dalton, & Grisham, 2007), and pronunciation (Cardoso, Collins, & White, 2012; Liakin, Cardoso, & Liakina, 2017; Soler-Urzua, 2011). Despite the positive evidence to support the use of TTS as a learning tool, the applications need to be formally evaluated for their potential to promote the conditions under which languages are acquired, particularly in an English as a *foreign* language (EFL) environment, as suggested by Cardoso, Smith, and Garcia Fuentes (2015).

The current study evaluated the voice of a modern English TTS system—used in an EFL context in Brazil—in terms of its speech quality, ability to be understood by L2 users, and potential for focus on specific language forms, and was operationalized based on the following criteria: (1) users' ratings of holistic features (comprehensibility, naturalness, and accuracy, as defined by Derwing & Munro, 2005); (2) intelligibility (the extent to which a message is actually understood), measured with a dictation task; (3) text comprehension (i.e., users' ability to understand a text and answer comprehension questions); and (4) users' ability to hear a specific morpho-phonological feature (i.e., the aural identification of English past tense -ed.)

Twenty-nine Brazilian EFL learners listened to stories and sentences, produced alternately by a TTS voice and a human, and rated them on a 6-point Likert scale according to the abovementioned holistic criteria (comprehensibility, naturalness, and accuracy). In addition, they were asked to answer a set of comprehension questions to assess their ability to understand what

they had heard. To measure intelligibility, participants completed a dictation task in which they were asked to transcribe utterances, as recommended by Derwing and Munro (2005). Finally, participants performed an aural identification of 16 sentences to judge whether the target feature (past mark -ed) was present or not. After these tasks were completed, semi-structured interviews were conducted to collect data regarding participants' perceptions of the technology.

Results indicate that the performance of both the TTS and human voices were perceived similarly in terms of comprehensibility, while ratings for naturalness were unfavorable for the TTS voice. In addition, participants performed relatively similarly in response to both voices with respect to the tasks involving text comprehension, dictation, and identifying a target linguistic form (past -ed) in aural input. These findings suggest that TTS systems have the potential to be used as pedagogical tools for L2 learning, particularly in an EFL setting where natural occurrence of the target language is limited or non-existent.

Acknowledgements

I would first like to thank my advisor, Dr. Walcir Cardoso, for guiding me since before the start of this program. His patience and dedication have been incommensurable. I would also like to thank my committee members, Dr. Laura Collins and Dr. Denis Liakin, for granting me some of their precious time, for providing invaluable feedback, and for agreeing to be part of my journey as a graduate student.

Dr. Walcir Cardoso recently reminded me of these wise words: "It takes a village to raise a child." Indeed, many people helped me to assemble this project. First, I would like to acknowledge Jennica Grimshaw, Ross Sundberg, and Kym Taylor for using their native-speaker magic to turn my scribbles into academic work. If they ever decide to write in Portuguese, I hope they will let me return the favor. I would also like to express my gratitude to Randall Halter, who helped me make sense of the statistics employed in this study. Finally, lots of *obrigados* and *gracias* go to Alexandre Bacelar and Moises Garibay for enduring my daily complaints and for motivating me to always move forward. I would not be strong enough without their precious support.

I would like to thank my dearest friends and family, specially Maria Luiza Bione for her assistance during the data collection. Above all, I would like to thank my mother, Maria José Bione, who never allowed me to stop dreaming, and my father Carlos José Alves da Silva, who celebrates my achievements from heaven.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

In any approach to second language (L2) acquisition, input is an essential component for learning (Gass & Mackey, 2007; Krashen, 1985), and language learners need to be exposed to a significant amount. Ellis (2002) argues that input frequency is intimately related to all aspects of language acquisition. For instance, there is conclusive evidence showing that chances for vocabulary acquisition increase when input provides learners with enough encounters with new words (Horst, Cobb, & Meara, 1998; Nation & Wang, 1999; Webb, 2007).

In addition to quantity, input quality is equally important for L2 acquisition (Ellis & Collins, 2009). According to the Common European Framework of Reference for Languages (Council of Europe, 2001), learners need to accrue more than 500 guided learning hours in order to reach a level of intermediate proficiency in a foreign language. Instruction is crucial, because L2 learners may not notice less salient linguistic forms to which they have access through naturally spoken or written language (Ellis, 2006). One of the alternatives to help students process positive evidence (VanPatten, 2007) is to manipulate instructional input in a way that increases the salience of opaque constructions to facilitate students' intake (Collins, Trofimovich, White, Cardoso, & Horst, 2009).

Input may be manipulated in numerous ways; some studies propose that input quality may be improved via an increase in variability. Barcroft and Sommers (2005) compared vocabulary gains between groups with dissimilar degrees of acoustic variability. Three groups learned new words in different settings: a) no variability, with only one speaker repeating each word six times; b) moderate variability, with three speakers repeating each word twice; and c) high variability, with six speakers repeating each word one time. Results showed positive effects of acoustic variation as higher-variability conditions improved vocabulary gains. The authors view these

results as support for the elaborative processing hypothesis, which proposes that memory traces are better retrieved after elaborate information processing (Lin, Fisher, Winstein, Wu, & Gordon, 2008). In this sense, Barcroft and Sommers concluded that "acoustically varied instances of each new lexical item in the input combine to form a representation that is more robust than would have been obtained by an equivalent number of acoustically consistent instances of the same item" (p. 405). The elaborative processing hypothesis is not restricted to vocabulary gains and may be extended to other linguistic skill acquisition, such as pronunciation. Therefore, one could hypothesize that acoustic variation in the input may also form better representations for L2 phonology, which is the focus of this study.

In sum, the current SLA literature recommends that an ideal learning environment should offer generous amounts of comprehensible input produced by variable sources, which should encourage researchers to study ways to provide EFL students with tools to increase their access to the language and to foster an autonomous learning style to overcome the input-related limitations regularly found in the L2 classroom. However, implementing these practices in a language classroom can be challenging due to the short amount of time available for students to be in contact with their target language in a formal instructional setting (Lightbown, 2003).

In an effort to find alternative ways to promote the ideal L2 learning setting, SLA research has turned its attention to Computer-Assisted Language Learning (CALL), a field that is well-represented by numerous organizations and publications (Levy & Hubbard, 2005). Several studies have investigated the effects of different CALL modalities, and among the plethora of available options, one type of technology has stood out for its natural capacity to offer extra language input both inside and outside the classroom: text-to-speech synthesizers, which are speech synthesis applications used to create spoken (oral) versions of textual input on personal

computers or mobile devices. Prior research has attested to the advantages of using TTS for developing different linguistic skills (Kirstein, 2006; Proctor, Dalton, & Grisham, 2007), including pronunciation (Cardoso, Collins, & White, 2012; Liakin, Cardoso, & Liakina, 2017; Soler-Urzua, 2011).

Despite positive evidence demonstrating the pedagogical benefits of Text-to-Speech synthesizers (TTS) for second/foreign language learning, there is a need for up-to-date formal evaluations, specifically regarding its potential to promote learning. This study evaluates the voice quality of a TTS system in comparison with a human voice, and examine its pedagogical potential for use in an English as a foreign language (EFL) setting in terms of its speech quality, ability to be understood by L2 users, and potential to focus on a specific language form. The following section describes the criteria under which previous studies have analyzed this technology thus far and establish the studies objectives.

Literature Review

An initial step for evaluating synthetic speech is to assess how it differs from natural speech. Researchers have drawn on previous studies of listeners' reactions to non-native speech to analyze TTS speech quality. For instance, evaluations of L2 speakers' pronunciation in general (Derwing & Munro, 2005) require the assessment of three aspects considered essential for communication among L2 users: (1) comprehensibility, or how difficult it is to understand an utterance, (2) intelligibility, or the extent to which a message is actually understood by interlocutors or listeners, and (3) accentedness, or how much an L2 accent differs from the L1, which includes the accent variation that characterizes native speech. Since synthetic voices may be programmed with any accent or voice features (e.g., voices that differ by gender or voice pitch), the concept of accentedness for TTS needs to be operationalized into three variables: (i)

naturalness, or how human-like a TTS voice sounds; (ii) pronunciation accuracy, or how well a TTS-produced voice emulates intelligible English phonological patterns, and (iii) acceptability, or how favorable a target voice is perceived by humans (see Cardoso et al., 2015, for a similar approach).

Evaluations of TTS systems over the past two decades have been limited (Bailly, 2003; Delogu, Conte, and Sementina, 1998; Kang, Kashigawi, Treviranus, and Kaburagi, 2008; Nusbaum, Francis, and Henly, 1995; Stevens, Lees, Vonwiller, and Burnham, 2005). The most common method has been to judge TTS and human speech samples using the set of categories mentioned above. However, based on the handful of studies available, previous research has not arrived at a consensus regarding the quality of TTS-produced voices compared to that of humans. What may explain differences in previous results is the use of inconsistent methods to assess TTS-generated voices. For example, previous studies have used different criteria in their evaluations, rather than taking a comprehensive, holistic view on the assessment of TTS-produced voice quality; in addition, most studies used native speakers' judgement for the evaluation of TTS and, therefore, might not be generalizable to second or foreign language speakers. Furthermore, investigations are relatively dated, having been conducted between eight and 23 years ago (but see Cardoso et al., 2015, for an exception in the context of *second* language users, as will be discussed below). Synthetic speech technology has improved considerably over the past two decades, particularly since the advent of voice-based personal assistants found in GPS systems, smartphones (e.g., Siri, Cortana), and speaking robots or personal assistants (e.g., Amazon Echo, Google Home). Finally, previous studies have failed to investigate a crucial element in evaluating the pedagogical effectiveness of any tool for L2 development, which is TTS's potential for affording students to focus on specific language features.

One exception to this, however, is a recent study by Cardoso et al. (2015) in which an evaluation of an up-to-date English TTS system was performed regarding its speech quality and potential to draw students' attention to linguistic forms. Moving beyond previous TTS studies, a new layer was added by evaluating the technology in terms of its potential to allow learners to focus on a linguistic feature, using a task that targeted the aural identification of English past tense *-ed* allomorphy: [t], [d], and [ɪd], as found in inflected past forms such as "walk[t]", "drag[d]" and "add[ɪd]", respectively. Results showed that the voices produced by the TTS were rated significantly lower than the human-produced samples for all four categories of speech quality (comprehensibility, naturalness, pronunciation accuracy, and intelligibility). However, excluding naturalness, TTS rating was still considered high (above 80% for comprehensibility, accuracy, and intelligibility). Regarding the potential for focus on a linguistic form, the TTS- and human-produced samples had similar results, indicating that, regardless of the source of delivery (human or TTS), participants were equally able to perceive the target past *-ed* allomorph (t, d, or ɪd) in short and decontextualized phrases (i.e., without temporal indicators such as "yesterday" and "last week"). The implication of this finding is that modern TTS systems are ready to be used for language learning activities, particularly as a supplemental source of input in terms of both quantity and quality. The authors concluded by suggesting directions for future research, in which they called for further studies involving *foreign* language contexts, particularly those in which opportunities for naturally-occurring English input are scarce or non-existent, similar to the environment observed in non-English-speaking countries such as Brazil. Thus, the goal of this quasi-replication study is to evaluate TTS synthesizers in an English as a *foreign* language (EFL) setting in Brazil, as will be discussed in detail in the next chapter.

Considerable dissimilarities in language exposure and learning settings may create distinctive demands from and for ESL and EFL students. Thus, it is hypothesized that a change in learning environment (from second to foreign) may positively affect learners' perceptions and attitudes towards TTS-produced input, as EFL students may perceive synthetic voices as an additional source of quality input, which is naturally lacking in their learning environment.

The objective of this study is to evaluate the voice quality of a TTS system in comparison with a human voice, and consequently examine its pedagogical potential for use in an EFL setting, following Cardoso et al.'s (2015) recommendation. As the supervisor of the work presented here, Cardoso has been involved in the conceptualization of the study as well as in the interpretation of findings. Because this is a manuscript-based master's thesis, Chapter 2 consists of a research paper ("a full submittable draft of a manuscript", as indicated in the MA thesis guidelines) in which parts of this chapter may be repeated in an expanded or abbreviated form.

Chapter 2

Second language (L2) researchers and practitioners have explored the pedagogical capabilities of text-to-speech (TTS) synthesizers—speech synthesis applications that create spoken versions of written text—for their potential to enhance the acquisition of writing (Kirstein, 2006), vocabulary and reading (Proctor, Dalton, & Grisham, 2007), and pronunciation (Cardoso, Collins, & White, 2012; Liakin, Cardoso, & Liakina, 2017; Soler-Urzua, 2011). Despite the positive evidence to support the use of TTS as a learning tool, the applications need to be formally evaluated for their potential to promote the conditions under which languages are acquired, particularly in an English as a *foreign* language (EFL) environment, as recommended by Cardoso, Smith, and Garcia Fuentes (2015).

This study evaluated a modern English TTS system in an EFL context in Brazil in terms of its speech quality, ability to be understood by L2 users, and potential for focus on specific language forms, operationalized according to the following criteria: (1) text comprehension (i.e., users' ability to understand a text and answer comprehension questions); (2) intelligibility (the extent to which a message is actually understood), measured by dictation-type task; (3) users' ratings of holistic pronunciation features (comprehensibility, naturalness, and accuracy), as defined by Derwing and Munro (2005); and (4) users' ability to hear and identify a specific morpho-phonological feature (i.e., the aural identification of English past tense *-ed)*, which is produced as [t], [d], and [ɪd] depending on the preceding environment—see Celce-Murcia, Brinton, and Goodwin, *Teaching pronunciation: A course book and reference guide* (2010) for details.

In order to contextualize the current study and define its scope and goals, the following section defines text-to-speech synthesis, examines the reported benefits of using TTS for

language learning (including the importance of input quantity, quality, and variability in L2 acquisition), discusses the inherent differences of second and foreign language learning settings, and reviews previous TTS system's evaluations, as well as the criteria under which previous studies have analyzed this technology thus far.

Literature Review

In recent years, SLA research has turned its attention to Computer-Assisted Language Learning (CALL), a field that is represented by numerous organizations and publications (Levy & Hubbard, 2005). In her book *English Language and Technology*, Chapelle (2003) argues that from both cognitive and social perspectives, CALL tasks can offer L2 learners opportunities to receive enhanced input as well as interact with and produce the target language, all of which are recognized as essential for language acquisition. Prior research has investigated the effects of different CALL modalities such as Computer Assisted Training (Neri, Cucchiarini, Strik, & Boves, 2002; Thomson, 2012), Computer-Mediated Communication (Díez-Bedmar & Pérez-Paredes, 2012; Fiori, 2005; Smith, 2004), Automatic Speech Recognition (Chiu, Liou, & Yeh, 2007; Liakin, Cardoso, & Liakina, 2015; Neri, Cucchiarini, & Strik, 2003), and Mobile Gaming (Grimshaw, Cardoso, & Waddington, 2016; Sundberg & Cardoso, 2016). Among the plethora of available options, one type of technology has stood out for its natural capacity to offer additional language input both inside and outside the classroom: text-to-speech synthesizers.

Defining Text-to-Speech Synthesizers

Text-to-speech (TTS) is a type of speech synthesis application that is used to create a spoken (oral) version of textual input on personal computers or mobile devices. Handley (2009) explains that "in very simple terms, speech synthesis is the process of making the computer talk" (p. 906). Indeed, most computers—such as Siri for Apple, Cortana for Windows, Alexa for Amazon's personal robot *Echo*, and the Google Translate App—now have the ability to "talk" via their built-in TTS features.

The Benefits of Using TTS for Second/Foreign Language Acquisition

Some studies attest to the advantages of using TTS for developing different linguistic skills. To examine how TTS could support L2 English learners' writing processes, Kirstein (2006) analyzed data from six high school students. The data consisted of essays (written with and without TTS support), questionnaires, documents, interviews, and observations. Findings suggested that when participants used TTS, they wrote more drafts, spent more time on each draft, and detected more errors. Related studies have also found that TTS is useful for vocabulary acquisition and reading training, as its read-aloud functionality reduces the decoding demands of many challenging texts (Proctor, Dalton, & Grisham, 2007; Rose & Dalton, 2002).

TTS seems to be particularly well-suited for pronunciation practice. Soler-Urzua (2011), for instance, designed an experiment to test the effects of TTS on phonological acquisition. She divided 47 Spanish-speaking participants into three instructional conditions: a) TTS-based instruction, b) non-TTS-based instruction, and c) regular classroom instruction (control group). All three groups were pre-tested for their ability to perceive and produce different vowel qualities (i.e., the distinction between /i/ and /ɪ/, the vowels in "beat" and "bit", respectively). After treatment, participants completed a post-test and a delayed post-test. Results showed that even

though the TTS group outperformed the non-TTS and control groups in perceiving and producing the English /i/-/ɪ/ contrast, the overall improvement in the TTS group was not significantly different from the non-TTS group. Nevertheless, the author observed a trend showing improvements in perception and production by the TTS group, a pattern that was not observed for the other two groups.

In order to justify the pedagogical usefulness of TTS in the classroom, however, positive effects are not enough; prior to implementation, any SLA material must be evaluated for its pedagogical usefulness through well-established theoretical frameworks to produce reliable and comparable results (Jamieson & Chapelle, 2010). Hence, TTS needs to be thoroughly examined under the light of relevant theory and research in SLA before being promoted as a pedagogical tool. However, what should researchers evaluate, and which measures should they use to evaluate TTS speech quality?

TTS Evaluation: Speech Quality

An initial step toward evaluating synthetic speech is to assess how it differs from natural speech. In other words, how does the quality of modern synthetic voices compare to human voices? To analyze TTS speech quality, researchers have drawn on previous studies of listeners' reactions to non-native speech. For instance, to evaluate L2 speakers' pronunciation in general, Derwing and Munro (2005) proposed a method that focuses on three aspects considered essential for communication among L2 users: (1) comprehensibility, or how difficult it is to understand an utterance, (2) intelligibility, or the extent to which a message is actually understood by an interlocutor or group of listeners, and (3) accentedness, or how much an L2 accent differs from the L1, including the variations in accents that characterize native speech.

In the context of synthetic voices, produced by software applications programmed with accent or language variations rather than by human speakers, the concept of accentedness may be viewed as consisting of three variables: (i) naturalness, or how human-like a TTS voice sounds, (ii) pronunciation accuracy, or how well a TTS-produced voice emulates intelligible English phonological patterns, and (iii) acceptability, or how favorable a target voice is perceived to be by humans (see Cardoso et al., 2015, for a similar approach). These concepts will be discussed later.

There have been a few evaluations of TTS systems and their voices over the past two decades. The favored method has been to judge TTS and human speech samples under the set of categories mentioned above. For example, in a study by Nusbaum, Francis, and Henly (1995), the authors compared TTS-produced voices in English to their human counterparts for naturalness in both segmental and suprasegmental features. In their first experiment, they instructed native English-speaking subjects to evaluate utterances of the segments /a/, /i/, and /u/ using a naturalness scale to measure the probability of a sound to be considered natural. Results differed between vowel categories, as TTS was perceived to be more natural than human voices for /a/, less natural for /i/, and equally natural for /u/. In a second experiment, L1 English participants evaluated prosody at the word level, also using a naturalness scale. The researchers manipulated the input to isolate prosody by removing all the segmental information from the stimuli. Therefore, participants were only able to listen to rhythmic word patterns produced by TTS and human voices. Their findings showed that even with the intelligibility variable removed, participants would still judge human voices to be more natural than TTS. Stevens, Lees, Vonwiller, and Burnham (2005) echoed these results when they found that their native English-speaking participants rated TTS sentences to be less natural than human-produced sentences. Other studies, though, have found more positive results for TTS voices regarding naturalness.

Kang, Kashigawi, Treviranus, and Kaburagi (2008) asked Japanese-speaking participants to rate English TTS and human input at word and sentence levels. They found that TTS voice was perceived to be as natural as human production, at least at the word level. These results are partially substantiated by Stern, Mullennix, and Yaroslavsky's (2006) findings, as they observed TTS messages to be perceived as favorable as those produced by humans.

Other TTS evaluations have focused on cognitive factors in synthetic voice comprehension. Delogu, Conte, and Sementina (1998) designed two experiments to compare comprehension of electronic and human voices. In their first experiment, participants were asked to identify target Italian words within a sentence so that the authors could measure the length of time needed to perform the task, which they assumed to be an index of intelligibility. For example, sentences in which participants took less time to identify target words were considered to be more intelligible. In the second experiment, participants listened to short paragraphs in Italian, then completed a multiple-choice comprehension test designed to objectively evaluate synthetic voices in terms of intelligibility (see Goldstein, 1995 and Nye, Ingemann, & Donald, 1975 for a similar approach). Multiple-choice questions are assumed to activate higher-order cognitive factors involved in speech recognition: namely, perception, memory, and attention. Delogu et al.'s experiments demonstrated that, in general, comprehending synthetic (non-human) voices is more demanding, as response duration for the former was higher and the degree of text comprehension was lower. Still, the authors indicated that the difficulty level decreased as the subjects had more exposure to synthetic voices. In another study that focused on measuring intelligibility using a French TTS system, Bailly (2003) noticed that participants performed better in shadowing tasks when they had human voice input instead of TTS-produced input. Interestingly, in a more recent study, Kang et al. (2008) found no significant difference between

human and TTS speech for text comprehension (i.e., the participants' ability to understand text produced by humans in comparison with TTS).

It seems clear, based on the handful of studies available, that previous research has yielded mixed results regarding the quality of TTS systems as compared to the human voice. One reason for this discrepancy is the use of inconsistent methods. For example, rather than taking a comprehensive, holistic view on the assessment of TTS-produced voice quality, previous studies have used different criteria in their evaluations; while some studies have focused exclusively on users' perceptions regarding the synthetic voice's naturalness, (e.g. Nusbaum et al., 1995; Stevens et al., 2005), others included only comprehension measures (e.g. Bailly, 2003; Delogu et al., 1998). In addition, most studies have used native speakers to evaluate TTS, which may have impacted their results and, therefore, those results might not be generalizable to second or foreign language speakers. Furthermore, those investigations are relatively dated, with the most recent being from 2009. Text-to-speech synthesis has evolved considerably over the past two decades, particularly since the advent of voice-based personal assistants found in GPS systems, smartphones (Siri, Cortana), and speaking robots (Amazon Echo, Google Home). Finally, previous studies have not investigated TTS's potential for focus on specific language forms, which is a crucial element in evaluating the effectiveness of any tool for L2 pedagogy.

One exception to this, however, is a recent study by Cardoso et al. (2015), in which an evaluation of an up-to-date English TTS system's speech quality and potential to draw students' attention to linguistic forms was performed. Moving beyond previous TTS studies, a new layer was added to evaluate the technology in terms of its potential to allow learners to focus on a linguistic form. The task targeted the aural identification of English past tense *-ed* allomorphy: [t], [d], and [ɪd], as found in inflected past forms such as "walk[t]", "drag[d]" and "add[ɪd]",

respectively. Fifty-six university-level students in Canada, an English as a *second* language environment, performed a series of tasks to evaluate a TTS system, in which they heard utterances alternately produced by TTS and human voices. Both native and second language speakers participated in this study. Results showed that the samples produced by the TTS system were rated significantly lower than the human-produced samples for all four categories of speech quality (comprehensibility, naturalness, pronunciation accuracy, and intelligibility). However, excluding naturalness, TTS rating was still considered high (above 80% for comprehensibility, accuracy, and intelligibility). Regarding the potential to focus on a linguistic form, the TTS and human-produced samples had similar results, indicating that regardless of the source of delivery (human or TTS), participants were equally able to perceive the target past *-ed* allomorph (/t/, /d/, or /ɪd/) in short decontextualized phrases (i.e., without temporal indicators such as "yesterday" and "last week"). The implication of this finding is that modern TTS systems are ready to be used for language learning activities, particularly as a supplemental source of input in terms of both quantity and quality. In their conclusion, the authors suggested directions for future research by calling for studies involving *foreign* language contexts, particularly those in which opportunities for naturally-occurring English input are scarce or non-existent. Thus, the goal of this quasi-replication study is to address this recommendation in an English as a *foreign* language (EFL) setting in Brazil, as will be discussed next.

Differences in Learning Contexts: ESL versus EFL

It is attested in the EFL literature that students may have low exposure to the target language, both within class and outside of it (Collins & Muñoz, 2016). Foreign language class time is often limited to few hours a week, which is not enough to provide students with the amount of input and practice necessary for mastering foreign language skills, which is assumed

to be approximately 10,000 hours of practice (Ericsson, Krampe, & Tesch–Römer, 1993). In Brazil, for instance, *Idiomas Sem Fronteiras* (Languages Without Borders), a Brazilian language learning program at the university level, offers four hours of EFL instruction per week in 16-, 32-, 48-, or 64-hour courses for low-income students (Idiomas Sem Fronteiras, 2017). In the public-school system, the scenario is even less ideal, as the quantity of L2 English exposure is reduced to two hours of instruction per week (British Council Brasil, 2015).

A limited number of instructional hours are also observed in other EFL settings such as in certain Asian (Lu, 2008) and Arabic-speaking countries (Derakhshan & Khodabakhshzadeh, 2011). Ortega (2013) estimates that whereas students in second language contexts may accrue 7,000 hours of L2 exposure in five years of contact with the target language (in a conservative projection of 4 hours of exposure a day), EFL students, on the other hand, may have as little as 540 hours of L2 exposure from instruction only in the same period (i.e., less than 10% of what is observed in Ortega's conservative estimates for second language contexts). Therefore, by having less exposure to their target language outside the classroom, EFL students tend to greatly rely on their teachers for L2 knowledge and input (Tanaka, 2009), which can create a teacher-centered environment that is not ideal for learning (Chapelle, 2001). This environment is particularly negative for pronunciation training, as exposure to L2 phonology is limited to one teacher who uses only one variety of English or accent. (See Thomson, 2011 for the rationale behind the recommendation of providing students with a learning environment in which the input is highly variable).

Aware of these limitations, one may conclude that considerable dissimilarities in language exposure and learning settings may create distinctive demands from and for ESL and EFL students. Thus, it is hypothesized that a change in learning environment (from second to foreign)

may positively affect learners' perceptions and attitudes towards TTS-produced speech, as EFL students may perceive synthetic voices as a useful source of additional input, given the often-limited exposure to the target language in their learning environment.

The objective of this study is to evaluate the quality of a TTS voice in comparison with a human voice, and consequently examine its pedagogical potential for use in an EFL setting, following Cardoso et al.'s (2015) recommendation. This study is guided by the following research question: What is the quality of speech produced by a TTS system in comparison with that of a human, based on the following six assessment measures:

1. text comprehension (one's ability to understand a short anecdote)
2. intelligibility (the extent to which a message is actually understood by an interlocutor or group of listeners)
3. comprehensibility (one's perception of how easy it is to understand a message)
4. naturalness (the extent to which a message deviates from sounding machine-made)
5. pronunciation accuracy (the extent to which a message deviates from fluent/native speaker norms)
6. form identification (the participant's ability to identify linguistic forms in speech: the identification of past -ed forms)

Methodology

Participants

Twenty-nine Brazilian EFL adult learners (M = 9, F = 20) at an intermediate level of proficiency, in Recife (Pernambuco, Brazil) participated in the study. Their ages ranged from 18 to 33 years old (M = 23.6, SD = 4.9), and all spoke Brazilian Portuguese as their first language (L1). Participant proficiency was determined based on a number of criteria: (1) placement at their

language institution; (2) the call for participants (which emphasized the target language proficiency: “participants who are at the intermediate level”); (3) their self-assessment in a background questionnaire (Appendix A); and finally, (4) the researcher’s overall perception of their skills (e.g., if they could not follow instructions in English or could not understand the written materials, the participants were not included in this study). The participant pool was comprised of English students from two different EFL schools (a post-secondary Professional School and a University Language Institute). All were either undergraduate students (n = 21) or holders of a bachelor’s degree (n = 8). They participated in this research as volunteers and, accordingly, did not receive any compensation.

Design

This study considered two independent variables—TTS and human voice—and measured their effect in three general variables: (a) intelligibility (including text comprehension), (b) learners’ ratings on holistic pronunciation measures (i.e., comprehensibility, naturalness, and pronunciation accuracy), and (c) opportunity to identify a grammatical form (past -ed). Past literature has presented some options for assessment of the variables that have been implemented in this study. For instance, Delogu, Conte, and Sementina (1998) preferred a text comprehension test to assess intelligibility. Derwing and Munro (2005), on the other hand, have recommended different measures to evaluate listeners’ reactions to non-native speech: To evaluate intelligibility, they proposed a transcription task (similar to a dictation activity). For comprehensibility and accentedness, they suggested a scalar judgment task (ratings) using Likert-scale items. As for the ability to focus on a linguistic feature, a previous study (Cardoso et al., 2015) used a form identification test to measure participants’ accuracy in recognizing English

past tense through aural input. Details about these tasks will be provided below, as they are relevant to the current research.

This study opted for the following design: The data were collected in a one-shot individual session wherein each participant completed a set of tasks designed to assess each criterion pertinent to evaluating the quality of TTS and human speech. For intelligibility, participants completed a dictation task during which they were asked to transcribe TTS- and human-based utterances on an answer sheet (Appendix B), as suggested by Derwing and Munro (2005). In addition, participants listened to two short anecdotes (or short stories) and answered six multiple-choice questions (Appendix C) covering each story's main points. Each set of comprehension questions was divided into five specific questions and one interpretation question. In order to evaluate pronunciation holistically, as suggested by Derwing and Munro, participants rated the quality of the speech that they heard based on three categories: comprehensibility, naturalness, and pronunciation accuracy, using a 6-point Likert scale (Appendix D). Participants rated not only the two anecdotes described above, but also 12 short sentences (e.g., The boy watched the clock ticking on the wall). The rationale for the inclusion of these short sentences was that they could yield different results due to the low cognitive load required for their processing, as the participants needed to concentrate solely on speech quality, not understanding (see forthcoming discussions). Finally, for the ability to focus on a linguistic form, as in Cardoso et al. (2015), participants performed an aural identification task for 16 sentences in which they judged whether the target feature (past tense *-ed*) appeared in the oral input they heard. Participants had to decide if the action took place in the past (e.g., I called my mother) or not (e.g., I visit my cousin Sam) and then check the corresponding form on the answer sheet (Appendix E). Note that sentences in the present tense were added to this task as distractors. At

the end of the session, participants were interviewed (in their native Portuguese) about their insights on the quality of the TTS-generated voices.

Stimuli

For all tasks, participants listened to speech samples that randomly alternated between TTS and human voices. The TTS voice was Julie (by NeoSpeech, available at http://neospeech.com), a female North American speaker whose voice was used for the synthesis of the target texts and sentences (see forthcoming material description). Human speech was produced by a female native-speaker of the same North American dialect with similar speech properties (mezzosoprano-like voice of a well-educated female adult) and no prior voice training, in an effort to emulate the type of voice that students naturally encounter in their language classroom (cf. Cardoso et al.'s, 2015 use of a professional voice coach). She recorded the same text and sentences as Julie and was instructed to match Julie's speed and intonation. Both human and TTS samples were converted into WAV audio format (Mono, 16bit, 44.1KHz), which were then embedded in Microsoft PowerPoint slides for presentation to the participants.

Materials

Both the stories (anecdotes) and sentences were adapted from materials produced by the ALERT research project (Collins, White, Horst, Trofimovich, & Cardoso, 2011). As alluded to earlier, the varied text length (i.e., longer stories vs. shorter sentences) was chosen to provide dissimilar cognitive conditions for participants. When compared to simple sentences, short stories contain more complex structures and may require more cognitive effort, which may impact intelligibility and participant ratings. Each short story (Appendix F) had approximately 230 words and lasted for approximately the same amount of time, regardless of voice type: 1min:43sec and 1min:22sec for Julie's output, and 1min:44sec and 1min:19sec for the human-

recorded text. The comprehension test (Appendix C) for each short story consisted of six multiple choice questions, each with one correct and three incorrect responses to choose from. The questions were divided into two types: specific (e.g., Why did the woman go to the store?) and general (e.g., What do you think probably happened after?). Participants could score from 0 to 6 points on each test involving text comprehension.

In addition to short stories, participants were exposed to 38 short sentences in total for the three remaining tasks (mean word count for each sentence was 9 words, SD = 3.7), corresponding to 2–3 seconds of speech for each. Sentence distribution among tasks is described below. For the intelligibility assessment, 10 sentences (Appendix G) were generated. In a task similar to dictation, participants heard each sentence (e.g., He saw a pregnant woman on the other side of the room) only once, after which they were asked to transcribe what they heard on an answer sheet. It was assumed that sentences that were more intelligible would yield a higher percentage of correctly transcribed words. Students' orthography inaccuracies were ignored, as the task was not intended to measure writing skills, but rather the extent to which participants could hear and comprehend English utterances. Participants could score 0–100% on each sentence, where 0 corresponded to no intelligibility at all and 100% represented the highest intelligibility level of any given utterance.

For the holistic assessment of pronunciation in terms of comprehensibility, naturalness, and accuracy, 12 sentences (Appendix H) were designed to match the vocabulary and morphosyntactic knowledge of intermediate-level learners so that the participants could focus exclusively on these three impressionist measures. In addition, the target sentences were constructed without any references to contextual cues (including references to past events).

After listening to each sentence, participants were asked to rank what they heard using a 6-point Likert scale, based on three questions: How easy was the voice to understand? (comprehensibility); How natural was the voice? (naturalness); and, How correct was the pronunciation? (accuracy). The term "correct" was chosen as a user-friendly term so that the novice listeners could understand the question. (Its use is based on feedback from a pilot test conducted with a small number of native and non-native English-speaking participants.) The category was described as how much the target voice deviated from a fluent/intelligible or native English speaker.

Finally, the linguistic feature identification material (past -ed) consisted of 16 sentences (Appendix I) carefully designed to avoid any lexical cues that could help participants to identify the tense without using morpho-phonological processing (e.g., words such as yesterday, usually, etc). This way, participants' judgments were taken based solely on their aural perception. After listening to each sentence once (e.g., I opened the door for her), participants were asked to decide whether the action took place in the present or past. Table 1 shows the distribution of present (conceived as distractors) and past sentences in the stimuli as well as the allomorphic distribution among the past sentences (note that the allomorphy is provided for illustrative purposes only, as its identification was not one of the targets of the current study).

Table 1

Distribution of present/past sentences and allomorphy

Tense	Total #	/d/	/t/	/ɪd/
Present (Non-past)	4	-	-	-
Past	12	3	4	5

The instruments used in this study, the aspects they are designed to test, the tasks in which they were included and how the data that they elicited were analyzed are summarized on table 2.

Table 2

Summary of instruments, measures, tasks, and analysis

Instrument	Measure(s)	Tasks	Analysis
Text comprehension (short stories)	Intelligibility	Comprehension test (n=2)	Average of correct answers
Dictation (sentences)	Intelligibility	Sentence transcription (n=10)	Percentage of transcribed words
Learners' ratings	Comprehensibility Naturalness Pronunciation Accuracy	Stories (n=2) and sentence (n=12) ratings	Average of ratings in a 6-point Likert scale
Aural identification of past tense -ed	Opportunity to identify a grammatical form	Tense identification	Percentage of correct identification

Procedure

To complete all tasks, participants had approximately one hour in one individual session (one-shot design). Before the session started, they were asked to read and sign a consent form, after which they received a brief description of the project and of the rating categories. However, it was not disclosed to the participants that they would listen to different voice types or that they would hear synthetic voices among the samples. Participants then proceeded with the Microsoft PowerPoint presentation to initiate the study. They listened to the target stimuli using headsets (Microsoft Lifechat LX-3000 Noise Cancelling Headset), wrote their answers, rated voices, and completed the dictation task on a printed answer sheet as they advanced task by task in the presentation.

The material presented to participants was organized in two randomized sequences (Sequence A and Sequence B) in a way that both sequences contained the same target sentences or texts, but were produced by different voice sources. For example, participants who received

Sequence A heard the same sentences as participants in Sequence B; however, all the sentences produced by TTS in Sequence A were recorded by human voice in Sequence B, and vice-versa.

At the end of the session, for completion and to assess participants' perceptions of the technology adopted, the researcher conducted a semi-structured interview in the participants' native language (Appendix J) to collect qualitative data about their perceptions and attitudes towards Julie, the TTS-produced voice adopted in the study. These qualitative data were used to enrich the discussion section, as they helped to understand some of the participants' answers and ratings. For the qualitative analysis, which is beyond the scope of the current study, see Bione, Grimshaw and Cardoso (2016).

Results

Participants' judgments of the stories and sentences (to measure comprehensibility, naturalness and accuracy), text comprehension results, percentage of correct words in their dictation task (to measure intelligibility), and their accuracy on identifying regular past (to measure TTS's ability to provide noticeable input) were tallied, and means of matched pairs were compared. Parametric statistics were used for data sets that meet the normality assumptions (namely data from short stories' text comprehension and ratings); for every other set, non-parametric tests were carried out. Paired sample t-test and Wilcoxon Signed-Rank tests were used respectively, with an alpha level of .05 for the determination of statistical significance. An adjusted alpha of .004 was calculated using a False Detection Rate (FDR) post-hoc method (Benjamini & Hochberg, 1995; Larson-Hall, 2010) to avoid false positive errors. As the Bonferroni adjustment may be too conservative when the number of comparisons is high (this study includes nine comparisons), which may lead to false negative errors, an FDR was deemed more suited for this analysis (see Herrington, 2002 for the rationale behind this decision).

Intelligibility

Intelligibility was measured at two cognitive levels: a text comprehension test for short stories (complex cognitive level) whose scores could vary from 0 to 6 on each story depending on how many questions were correctly answered, and sentence transcription for sentences (simple cognitive level), where participants could transcribe between 0% to 100% of each sentence depending on the number of words correctly transcribed. The details for each analysis are described below.

Short stories (text comprehension).

A Paired sample t-test was conducted to compare how intelligible TTS- and human-narrated short stories were. There was no significant difference in the scores for TTS (M = 4.57, SD = .81) and human (M = 4.74, SD = .75); t(1) = -4.25, p = .147. Figure 1 illustrates the results for each story. These results suggest that the type of voice input that the participants received to complete the listening comprehension task had no impact on intelligibility for either story.

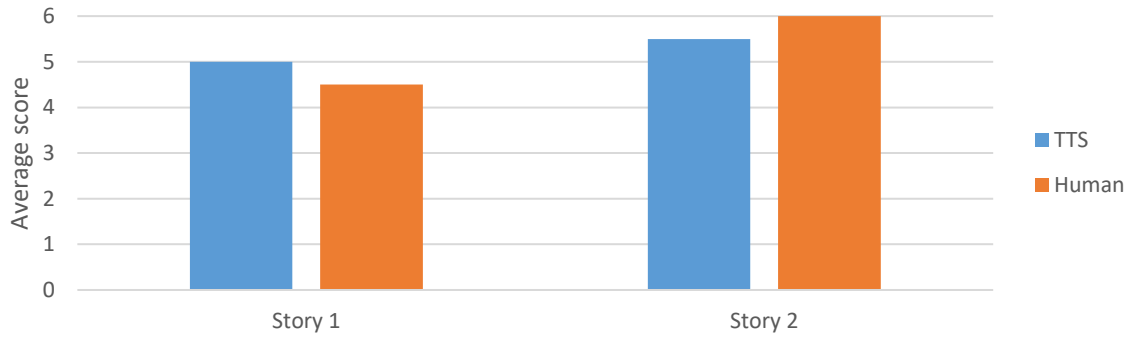

Figure 1. Short stories' score average

Sentences (dictation task).

A Wilcoxon Signed-Rank test was conducted to compare intelligibility in sentences produced by either TTS or human voice. There was no significant difference in the scores for TTS (Mdn = 59.65%) and human samples (Mdn = 55.05%); Z = -.153, p = .878. As roughly 60%

of all words within the sentences were transcribed, regardless of their source, these results show that the type of voice did not affect intelligibility at simple cognitive levels. Figure 2 shows the distribution of correctly transcribed words for each sentence pair by all participants.

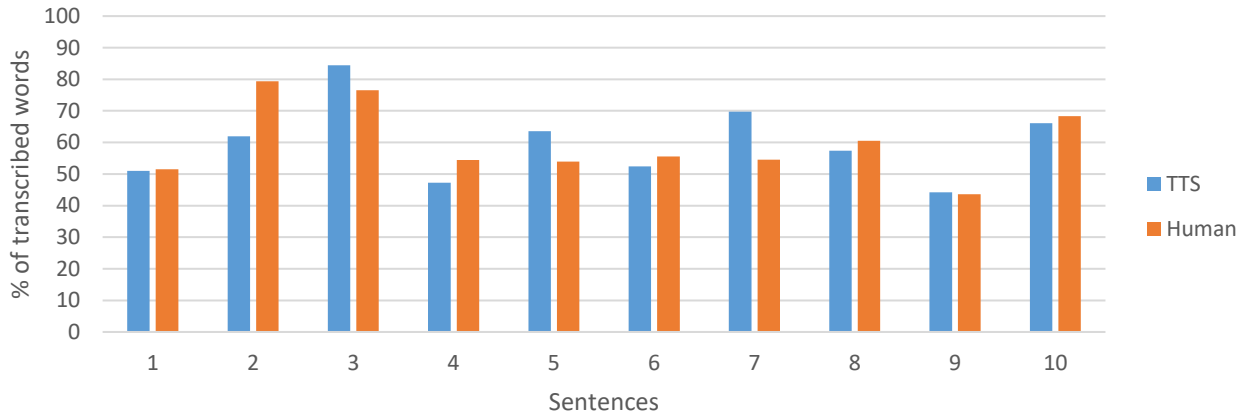

Figure 2. Percentage of transcribed words by sentence

Users' Ratings: Comprehensibility, Naturalness, and Accuracy

For each aspect under users' ratings, paired sample t-test (for short stories' ratings) or Wilcoxon Signed-Rank tests (for sentence ratings) were conducted. Statistical test results are reported below.

Short stories.

Considering an adjusted alpha of .004, paired sample t-tests yielded no significant difference in ratings for any aspect, as shown in Table 2. These results indicate that when listening to short stories, participants did not find substantial dissimilarities between samples.

Table 3

Short story holistic ratings

Aspect	TTS		Human		t	p
	M/6	SD	M/6	SD		
Comprehensibility	4.42	.02	4.92	.30	-2.59	.235
Naturalness	3.12	.74	4.58	.41	-6.35	.099
Accuracy	5.04	.15	5.31	.13	-27.00	.024

Sentences.

Based on Wilcoxon Signed-Rank tests, human samples were considered significantly more natural and more accurate than TTS samples. On the other hand, no significant difference was found for comprehensibility. Table 3 summarizes these results.

Table 4

Sentence holistic ratings

Aspect	TTS	Human	Z	p
	Mdn/6	Mdn/6		
Comprehensibility	5.06	5.10	-.628	.530
Naturalness	3.45	5.13	-3.06	.002*
Accuracy	4.93	5.10	-2.85	.004*

*$p < .004$

For a more detailed illustration of the results presented in Table 3, figures 3, 4 and 5 provide the distribution of user ratings for each sentence used in the respective test.

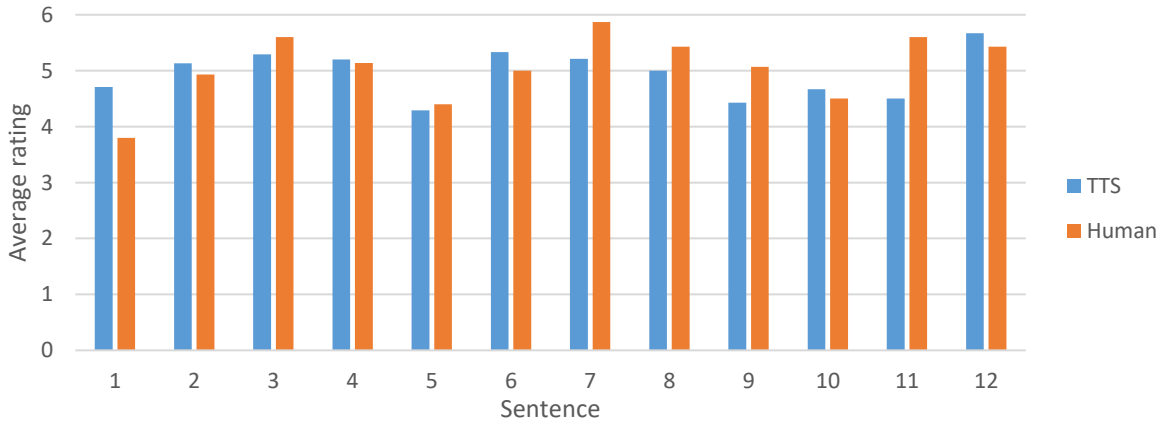

Figure 3. Comprehensibility rating distribution across 12 target sentences

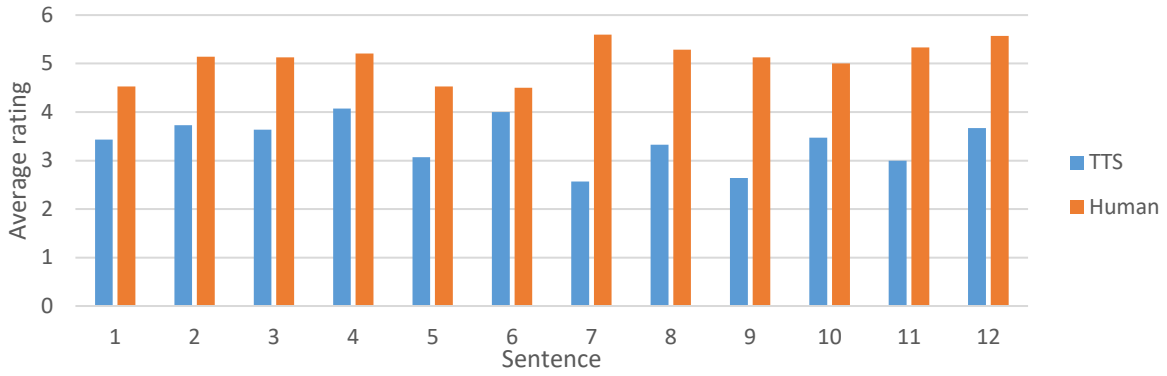

Figure 4. Naturalness rating distribution across 12 target sentences

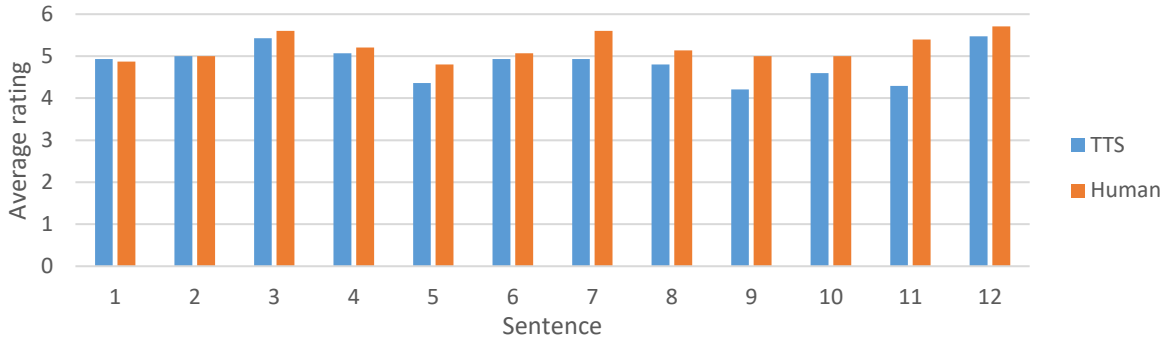

Figure 5. Accuracy rating distribution across 12 target sentences

These results indicate that text complexity may affect students' ratings, since TTS was rated as natural and accurate as human voice in the presence of cognitively complex input (short stories), but received significantly lower ratings for those two aspects when cognitive complexity decreased (sentences). Conversely, comprehensibility seems unaffected by text complexity, as TTS and human voice were equally comprehensible for participants at both simple (sentence rating) and complex input levels (story rating).

Aural Identification of a Linguistic Form (Past -ed)

A Wilcoxon Signed-Rank test showed no significant difference in answer accuracy between voice types. In other words, participants were equally able to recognize if a sentence was set in the simple past for both TTS (Mdn = 76%) and human samples (Mdn = 85%); Z = -1.735,

p = .083. Figure 6 displays the percentage of correct identification by voice type for each past tense sentence.

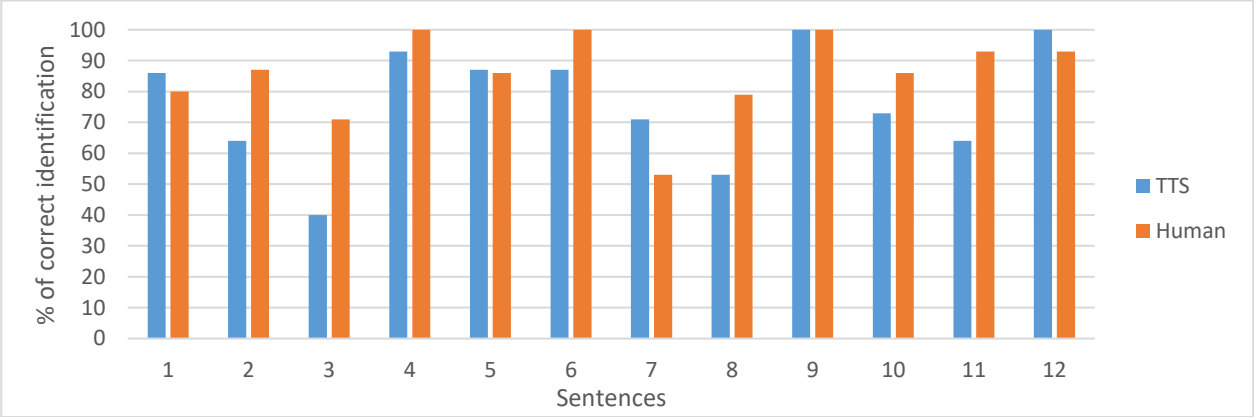

Figure 6. *Score of aural identification by past tense sentence*

Participants seem to behave similarly with the distractors (present tense) since the data did not show a noticeable difference between voice types. For a comprehensive distribution of results regarding the participants' ability to identify both past and present forms in the target voices as well as the representation of past tense allomorphy in the sentences, see figure 7.

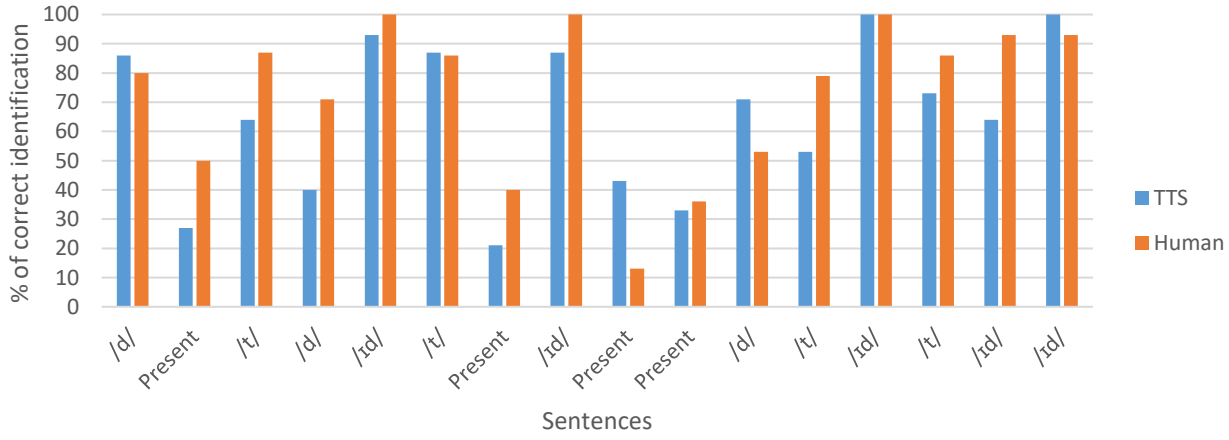

Figure 7. *Score of aural identification by sentence and distribution of past sentences, distractors (present), and allomorphy*

Discussion

This study evaluated the voice quality of a TTS system in comparison with a human voice, and consequently examined its pedagogical potential for use in an English as a foreign language setting. The following research question was addressed: What is the quality of speech produced by a TTS system in comparison with that of a human? The answer to this question was based on six assessment measures: text comprehension (one's ability to understand a short anecdote), intelligibility (the extent to which a message is actually understood by interlocutors or listeners), comprehensibility (one's perception of how easy it is to understand a message), naturalness (the extent to which a message deviates from "machine-made"), pronunciation accuracy (the extent to which a message deviates from fluent/native speaker norms), and opportunities for grammatical feature identification (one's ability to identify regular past tense). These measures encompass three general aspects of L2 pronunciation assessment: a) intelligibility (at two distinct cognitive levels: complex short stories and simple short sentences), b) users' holistic ratings (including comprehensibility, naturalness and pronunciation accuracy), and c) opportunity for focus on a linguistic form (past -ed).

Analysis of the data collected in the study showed that EFL learners rated or performed similarly, regardless of the input source, except for the naturalness and accuracy aspects at the sentence level only (not in longer narratives). Overall, these results correspond to what Kang et al. (2008) found in their research involving non-native English speakers, wherein they concluded that L2 learners do not recognize a remarkable difference between synthetic and human voices. A discussion of the results obtained for each feature under investigation is provided below.

Intelligibility: Text Comprehension and Dictation Task.

Previous studies have most commonly reported that TTS presents low intelligibility when compared to natural speech. For instance, Delogu et al. (1998) concluded that the user's cognitive load is heavier in synthetic voices because listening to TTS is a more demanding task than listening to humans, possibly due to the unexpected pauses and/or other prosodic limitations observed in synthesized voices. Bailly (2003) presented similar results, as his participants performed better in shadowing involving human voices than those using artificial voice samples. Contrary to previous studies where TTS scored lower than human voice, the current research revealed that both voice sources were equally intelligible. This contrast with previous results may be due to two factors: the new advances in TTS technology and the participants' increased exposure to electronic voices, as will be discussed next.

Elaborating upon TTS's previously-reported poor results, Bailly (2003) suggests that they were mainly due to the inappropriate prosody generated by the technology available at the time. It is out of the scope of this work to compare current and previous versions of TTS applications, but if we consider that almost 15 years have passed since Bailly's experiment, we may comfortably assume that speech technology has advanced considerably. As indicated by Handley (2009), current text-to-speech systems have not yet reached an optimal development stage at the prosodic level; however, the data presented in this study show that they have at least evolved to the point where their voice quality does not affect intelligibility (see forthcoming discussion on learners' perceptions of TTS prosody).

Regarding the hypothesis that an increase in exposure to electronic voices may lead to a higher acceptability, Delogu et al. (1998) noticed that intelligibility increased when participants became more acquainted with synthetic voices. If Delogu et al.'s remarks about a positive

correlation between exposure to electronic voices and intelligibility are accurate, then increasing access to these types of voices may explain why this study found no difference in intelligibility between synthetic and human voices. Since the boom of Apple's synthetic voice, Siri, in 2011, most commonly used computers and mobile devices offer built-in text-to-speech capabilities and, as a consequence, people have continuous access to artificial speech through GPS systems and their smartphones (e.g., Siri, Cortana). In addition, it is virtually impossible to contact any service provider without first interacting with an electronic voice that guides customers through menus before a human agent is reached. Although the current study did not measure participants' previous experience with these types of synthetic voices, we can ascertain that, due to their age (young and educated adults) and the ubiquitous use of synthetic voices in phone-based customer service, they are regularly exposed to TTS-generated voices.

Learners' Ratings of Holistic Pronunciation Measures: Comprehensibility, Naturalness, and Accuracy

The results involving users' ratings revealed that learners' judgement of TTS may be affected by the context in which the voices were used. For instance, participants rated TTS comprehensibility, naturalness, and accuracy as equal to the human voice when the task required more than simply emitting an opinion on each category (i.e., understanding a passage to answer a comprehension test and rating the related voices in the text). Participants clearly became more demanding when they were asked to focus exclusively on shorter oral texts (sentences). It was only in this context that they found that TTS sounded less natural or less accurate than human speech samples. These findings corroborate those found in previous research (e.g., Cardoso et al., 2015; Kang et al., 2008; Nusbaum et al., 1995; Stevens et al. (2005).

This difference in judgement may be explained by humans' limited processing capacity. Among several cognitive factors involved in processing a foreign language (e.g., perception, memory), attention plays a fundamental role (Schmidt, 1990). Since attention is a limited cognitive resource that permits subjects to focus their mental capacity on individual items (Delogu et al., 1998), cognitively demanding contexts may force attention away from peripheral information (in this case, perceptions of naturalness and accuracy) in order to process the content information conveyed in the speech. In this sense, participants may have shifted their attention to the text content so that they could comprehend the stories, thereby blurring any existing distinctions between TTS and human voices. When the cognitive load was lower, as with the sentence ratings, they attended to those distinctions more clearly and, consequently, they fine-tuned their speech perceptions. However, what exactly did they notice? Why was the synthetic voice judged to be less natural? Which aspects of human voice were inaccurately emulated by the TTS-generated voice?

Obviously, these research questions are beyond the scope of this work and, as mentioned earlier, the qualitative results are analyzed in Bione et al. (2016). Bione et al. show that when questioned about their opinions regarding the samples, students mostly complained about prosody. For example, one participant stated: "It was easy to understand, but I don't think it was correct. […] it doesn't have the right tone for pauses and commas." Another participant thought that "sometimes it spoke without pauses, sometimes it spoke slowly, word by word," and "speed and intonation made it sound like it could be a different word." In other words, suprasegmental characteristics such as pause, speed and intonation may have resulted in lower TTS ratings. Another cause may be related to the TTS "accent", as some participants did not notice that a synthetic voice was included in the testing samples: "It completely fooled me," "No, I didn't

realize that," or "No. Was there a computer voice?" said some participants after being informed that one of the voices was machine-produced. For those participants, the perceived distinctions were probably similar to those that characterize different human accents. As they did not recognize TTS samples as artificial, some participants believed that they heard Spanish, French or Indian accents. For instance, a participant said: "Since it sounded like a French speaker [speaking English], it had more intonation, sometimes resulting in a pronunciation that wasn't very natural."

Finally, for the last rating category, the results show that TTS and human voice were judged equally comprehensible for both short stories and sentences. These results do not support the findings reported in Cardoso et al. (2015), who found that the samples produced by the TTS system were rated significantly lower than those that were human-produced. This finding confirms the hypothesis that a change in learning environment (from second to foreign) could positively affect learners' perceptions and attitudes towards TTS-produced input, and suggests that *EFL* learners are less sensitive to distinctions between natural and artificial voice than *ESL* students. Low exposure to the target language and the resulting lack of L2 input in the foreign language environment may explain this difference, because when compared to ESL learners, EFL students have fewer opportunities to create strong and more accurate phonological representations of the L2.

Potential for Focus on a Linguistic Feature

The synthetic voice used in this evaluation was also able to match the natural voice in an identification task involving a morpho-phonological feature: the pronunciation of past -ed. No difference between voice sources was found in recognizing the presence of past tense morpho-phonology. As such, these findings corroborate those found in Cardoso et al. (2015) regarding the

opportunities afforded by TTS voices for students to notice distinctions in L2 input. These results may be explained by Julie's (the TTS voice) accuracy in reproducing English morpho-phonological patterns, as observed in a recent study by John and Cardoso (2016), in which the authors carried out a systematic evaluation of segmental and prosodic features of TTS and human output in order to establish the phonetic accuracy of the synthetic voice. In their evaluation (based on purely phonetic comparisons conducted by the researchers), problematic features of English phonology were targeted, including the TTS's ability to accurately reproduce past -ed allomorphy. Their results suggest that TTS performs equally to humans in pronouncing -ed forms and, in some contexts (e.g., producing the allomorph /d/), may even surpass humans. Based on our findings, supported by John and Cardoso's research, we may conclude that TTS-generated voices' ability to enhance the input for the noticing of past tense marking is similar to that of humans.

Conclusion

This study sought to evaluate the voice of a modern TTS in an English as a *foreign* language environment based on a set of assessment measures. It found that TTS-generated samples were comparable to human voice with respect to intelligibility, comprehensibility and ability to provide learners with opportunity to notice linguistic forms (similar to what human speech is capable of). On the other hand, the participants considered TTS-based voice less natural and less accurate when compared to the human voice in the context of short sentences.

The low ratings for these two aspects may appear negative, but based on the participants' insights during the interviews, this had little impact on their perception of TTS as a pedagogical tool for their own L2 learning. Most participants (23 out of 29) believed that synthetic voices should be used as learning tools (e.g., "If people start studying with computer voice, they'd get

used to it).” One student thought that TTS had the potential to help them establish a clearer relationship between phonology and orthography: “At the beginning, you relate sound to orthography, so you have to understand, especially the past marks. You have to reinforce it”.

The results obtained suggest that synthetic voices have the potential to deliver intelligible and comprehensible input, similar to human speech. From a pedagogical standpoint, this is beneficial because their use (preferably using a TTS application) can extend the reach of language classrooms by allowing students to practice on their own time and in their own space; more importantly, TTS may enhance (in both quantity and quality) learners’ access to the target language. In sum, the usage of TTS may provide a level-appropriate, user-controlled solution that produces accurate speech models for pronunciation practice and for the development of language awareness (e.g., to raise students’ awareness about the different realizations of the past -ed inflection), and thus assist in the acquisition of L2 morpho-phonological patterns.

There were several methodological limitations to the study. First, the small number of participants may prevent more assertive conclusions. Moreover, this study only considered intermediate English proficiency and, accordingly, is not able to determine whether this variable affected the results. Additionally, the high number of comparisons may have decreased statistical power; however, most results would remain unchanged even if an alpha level of .05 for statistical significance had been used (i.e., if the number of comparisons were fewer). Finally, due to the number of tests carried out during the experiment and the time limitations of a one-shot study, this research opted for a reduced quantity of tokens for some tasks (e.g., the past -ed feature identification task) so as to not overextend the session time or fatigue the participants.

For future voice quality evaluations, the investigation should consider a larger number of participants from different proficiency levels. It would also be wise to divide the experiment into

multiple sections with pauses in between so that the number of tokens may be increased without causing participant fatigue. Future studies should also evaluate CALL software using actual TTS applications for language learning: Would the results be different if the participants had access to all features available for TTS in which they can repeat forms at will and manipulate the input in terms of speed, pitch, or regional accent? Finally, to gather empirical evidence of TTS's potential as a pedagogical tool is to examine whether its use leads to learning gains (e.g., if its use facilitates the acquisition of regular past tense allomorphy), over an extended period of usage.

From a pedagogical perspective, Leow (2015) believes that it is the learners' responsibility to learn (as no one can learn for them) and to come to class prepared to practice, whereas teachers should offer students well-designed tasks to maximize their learning. In this context, TTS may help teachers develop suitable and personalized learning tasks for their students and have the potential to enhance the L2 learning environment by affording students the opportunity to select their own materials and, consequently, have an active role in the learning process.

Chapter 3

This chapter first expands upon the conclusions drawn in the previous section and examines some related phenomena that emerged during analysis. It will then contextualize the present evaluation within a broader CALL evaluation framework and discuss future directions for research.

General Conclusion

This study evaluated the voice quality of a TTS system in comparison with a human voice and, consequently, examined its pedagogical potential for use in an EFL setting. Not only did the findings show that TTS and human voice samples were comparable in most aspects of the assessment, they also confirmed that participants had an overall positive impression of TTS-generated voices. On the other hand, in agreement with previous studies (e.g., Cardoso et al., 2015; Kang et al., 2008; Nusbaum et al., 1995; Stevens et al., 2005), the TTS voice was rated less favorably in terms of naturalness and accuracy when compared to a human voice.

While the low ratings for the two aforementioned aspects may appear negative, they had little impact on participants' perception of TTS as a pedagogical tool. Almost all participants recognized that TTS could and should be used for teaching purposes, and most said that it should be used regardless of students' proficiency levels. This contrasts with Cardoso et al.'s (2015) findings, wherein participants showed lower acceptance towards the pedagogical use of TTS in a *second* language context. One reason for this high acceptance of TTS as a pedagogical tool in the current study may be due to the fact that EFL environments lack naturally-occurring L2 input and access to native or proficient speakers in the target L2. These findings suggest that EFL students, at least those included in this study, appear to be ready to adopt TTS systems as pedagogical tools in L2 education and endorse Cardoso et al.'s (2015) conclusion that "modern TTS systems seem

to be ready for advancement to further stages of evaluation, but more importantly, for use in language learning activities, particularly as a supplemental source of input which can cater to learners' individual needs and interests" (p. 112). The next section describes additional stages of evaluation that may be considered in future research.

Future Directions

Jamieson and Chapelle (2010) advocate that prior to classroom implementation, any CALL material must be evaluated for pedagogical purposes through recognized frameworks in order to produce stable, comparable, and defensible results. Thus, TTS as a CALL tool needs to be thoroughly examined under the light of relevant theory and research in SLA before it is deemed appropriate for adoption as a pedagogical tool. Chapelle (2001a, 2001b) proposed a three-stage framework to evaluate CALL applications, and it includes: (1) potential to provide ideal conditions to promote SLA, (2) analyses of activities using CALL software, and (3) empirical evaluation of learners' performance in such activities.

Regarding the first stage, Chapelle (2001b) established a set of criteria to evaluate the pedagogical potential of CALL tools. Table 4 summarizes her criteria and describes how the pedagogical use of TTS fits in each category. The current study addresses two aspects in Chapelle's framework, namely *language learning potential* and, to a lesser extent, *positive impact* (based on learners' attitudes towards its pedagogical use), but more evidence is required to observe how TTS corresponds to all criteria.

Table 5

Chapelle`s criteria for CALL evaluation (first stage)

Criteria	Description	TTS
Language learning potential	The degree of opportunity present for beneficial focus on form	TTS voices may offer opportunity for noticing forms that are not transparent in the input (e.g., past -ed, tense vs. lax vowel contrast as in "beat" and "bit").
Learner fit	The amount of opportunity for engagement with language under appropriate conditions given learner characteristics	Level appropriate, accessibility for learners, personalization
Meaning for focus	The extent to which learners' attention is directed toward the meaning of the language	The application of level-appropriate, authentic texts through TTS may facilitate the meaningful use of the language.
Authenticity	The degree of correspondence between the learning activity and target language activities of interest to learners out of the classroom	Use of authentic texts from real life, the internet, newspaper articles, learner-selected material, etc.
Positive impact	The positive effects of the CALL activity on those who participate in it	Previous studies found that TTS enhances the acquisition of writing, vocabulary, reading, and pronunciation; this study showed an overall positive attitude towards TTS as a pedagogical tool.
Practicality	The adequacy of resources to support the use of the CALL activity	Widely used technology, easily accessible, built-in feature in most computer and mobile devices

As for the second stage, effort on material development using synthetic voices is required. Handley and Hamel (2005) support that due to its flexibility and easy access, synthetic voices have the potential to be used in pedagogical activities. In addition, they have low storage requirements, the ability to generate speech models on demand, an ease of creation and

modification of exercises, and suitability for pronunciation training in both segmental and supra-segmental levels. Accordingly, this technology can be used in the development of activities that would include, for instance: (a) talking dictionaries that provide instant pronunciation models to help graphic-phonic form mapping, (b) talking texts to support reading comprehension activities, or (c) dictation tasks where learners can select the voice, style, speech rate and pitch. Interestingly, CALL software using TTS in such manners is already available and could form the base for the second evaluation stage (e.g., Rosetta Stone, LingQ, TinyCards).

At this stage, it would also be interesting to evaluate teachers' perceptions of using TTS as a learning tool. Research has shown that a successful integration of learning technologies into classrooms requires complex interactions between teachers, students, and technology (Cope & Ward, 2002; Honey, Culp, & Carrigg, 2000). Future research should focus on teachers' attitudes and personal beliefs towards the use of artificial voice for teaching and learning and try to answer questions such as: "Are teachers interested in using TTS as a pedagogical tool?", "If they are, do they feel ready to integrate this technology in the classroom?", and "What professional development is required for teachers to adopt synthetic voices in their classroom?".

Finally, for the last stage, empirical research needs to be carried out to attest learner's actual gain using TTS. Some effort has been already made in this sense. For instance, Liakin, Cardoso, and Liakina (2017) tested the impact of using mobile TTS on the L2 acquisition of French liaison, and they found that both control and experimental groups improved in liaison production, but when considered separately, only the experimental groups improved over time. Future research should confirm these results by evaluating TTS in EFL settings to verify if students in these contexts could also benefit from its adoption.

Concluding Remarks

Synthetic speech is no longer perceived as robot-like and, according to our findings, it has attained quality levels similar to human speech in terms of intelligibility and comprehensibility. In addition, TTS is a readily accessible technology and, due to its flexibility (users may adapt its voice, style, speech rate and pitch), it is a perfect candidate to be explored as a CALL tool. As such, it is not surprising that the technology has already started being used as a pedagogical tool to fulfill L2 learners' needs and, as a result, to help the paradigm shift from a teacher-centered to a more learner-centered environment.

References

Bailly, G. (2003). Close shadowing natural versus synthetic speech. *International Journal of Speech Technology, 6*(1), 11–19.

Barcroft, J., & Sommers, M. S. (2005). Effects of acoustic variability on second language vocabulary learning. *Studies in Second Language Acquisition, 27*(3), 387–414.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, 57*(1), 289–300.

Bione, T., Grimshaw, J., & Cardoso, W. (2016). An evaluation of text-to-speech synthesizers in the foreign language classroom: learners' perceptions. In S. Papadima-Sophocleous, L. Bradley & S. Thouësny (Eds), *CALL communities and culture – short papers from EUROCALL 2016, Limassol, Cyprus* (pp. 50-54). Dublin, IE: Research-publishing.net.

British Council Brasil (2015). *O Ensino de Inglês na Educação Pública Brasileira.* São Paulo, BR: British Council.

Cardoso, W., Smith, G., & Garcia Fuentes, C. (2015). Evaluating text-to-speech synthesizers. In F. Helm, L. Bradley, M. Guarda, & S. Thouësny (Eds.), *Critical CALL – Proceedings of the 2015 EUROCALL Conference, Padova, Italy* (pp. 108–113). Dublin, IE: Research-publishing.net.

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

Chapelle, C. A. (2001a). *Computer applications in second language acquisition: Foundations for teaching testing and research.* Cambridge, UK: Cambridge University Press.

Chapelle, C. A. (2001b). Innovative language learning: Achieving the vision. *ReCALL, 23*(10), 3–14

Chapelle, C. A. (2003). *English language learning and technology: Lectures on Applied Linguistics in the age of information and communication.* Amsterdam, NL: John Benjamins.

Chiu, T. L., Liou, H. C., & Yeh, Y. (2007). A study of web-based oral activities enhanced by automatic speech recognition for EFL college learning. *Computer Assisted Language Learning, 20*(3), 209–233.

Collins, L., Trofimovich, P., White, J., Cardoso, W., & Horst, M. (2009). Some input on the easy/difficult grammar question: An empirical study. *The Modern Language Journal, 93*(3), 336–353.

Collins, L., & Muñoz, C. (2016). The foreign language classroom: Current perspectives and future considerations. *The Modern Language Journal, 100*(S1), 133–147.

Cope, C., & Ward, P. (2002). Integrating learning technology into classrooms: The importance of teachers' perceptions. *Educational Technology & Society, 5*(1), 67–74.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge, UK: Cambridge University Press.

Delogu, C., Conte, S., & Sementina, C. (1998). Cognitive factors in the evaluation of synthetic speech. *Speech Communication, 24*(2), 153–168.

Derakhshan, A. & Khodabakhshzadeh, H. (2011). Why CALL why not MALL: An in-depth review of text-message vocabulary learning. *Theory and Practice in Language Studies, 1*(9), 1150–1159.

Derwing, T. M. & Munro, M. J. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly, 39*(3), 379–397

Díez-Bedmar, M. B. & Pérez-Paredes, P. (2012). The types and effects of peer native speakers' feedback on CMC. *Language Learning & Technology, 16*(1), 62–90.

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition, 24*(2)*,* 143–188.

Ellis, N. C. (2006). Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics, 27*(2), 164–194.

Ellis, N. & Collins, L. (2009). Input and Second Language Acquisition: The Roles of Frequency, Form, and Function Introduction to the Special Issue. *The Modern Language Journal, 93*(3), 329–336.

Ericsson, K. A., Krampe, R. T., & Tesch–Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*(3), 363–406.

Fiori, M. L. (2005). The development of grammatical competence through synchronous computer-mediated communication*. CALICO Journal, 22*(3), 567–602.

Gass, S. M. & Mackey, A. (2007). Input, interaction, and output in second language acquisition. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 175–199). New Jersey: Lawrence Erlbaum

Goldstein, M. (1995). Classification of methods used for assessment of text-to-speech systems according to the demands placed on the listener. *Speech Communication, 16*(3), 225–244.

Grimshaw, J., Cardoso, W., & Waddington, D. (2016). Can a 'shouting' digital game help learners develop oral fluency in a second language? In S. Papadima-Sophocleous, L. Bradley & S. Thouësny (Eds.), *CALL communities and culture – short papers from EUROCALL 2016, Limassol, Cyprus* (pp. 172–177). Dublin, IE: Research-publishing.net.

Handley, Z. & Hamel, M-J. (2005). Establishing a methodology for benchmarking speech synthesis for computer-assisted language learning (CALL). *Language Learning & Technology, 9*(3), 99–120.

Handley, Z. (2009). Is text-to-speech synthesis ready for use in computer-assisted language learning? *Speech Communication, 51*(10), 906–919.

Herrington, R. (2002). *Controlling the false discovery rate in multiple hypothesis testing.* Research and Statistical Support [Website]. University of North Texas, Denton, USA. Retrieved from https://it.unt.edu/sites/default/files/rss-id-false-discovery-rate-hypothesis-testing.pdf

Honey, M., Culp, K. M., & Carrigg, F. (2000). Perspectives on technology and education research: Lessons from the past and present. *Journal of Educational Computing Research, 23*(1), 5–14.

Horst, M., Cobb, T., & Meara, P. (1998). Beyond a clockwork orange: Acquiring second language vocabulary through reading. *Reading in a Foreign Language, 11*(2), 207–223.

Idiomas Sem Fronteiras (2017, January 28). Inglês Sem Fronteiras - Curso presencial [Website] Retrieved from http://isf.mec.gov.br/ingles/pt-br/curso-presencial

Jamieson, J. & Chapelle, C. A. (2010). Evaluating CALL use across multiple contexts. *System, 38*(3), 357–369.

John, P. & Cardoso, W. (2016). A comparative study of text-to-speech and native speaker output. In J. Demperio, E. Rosales & S. Springer (Eds.), *Proceedings of the Meeting on English Language Teaching* (pp. 78–96). Québec, CA: Université du Québec à Montréal Press.

Kang, M., Kashiwagi, H., Treviranus, J., & Kaburagi, M. (2008). Synthetic speech in foreign language learning: An evaluation by learners. *International Journal of Speech Technology, 11*(2), 97–106.

Kirstein, M. (2006). *Universalizing universal design: Applying text-to-speech technology to English language learners' process writing* (Doctoral dissertation). University of Massachusetts, Boston, MA.

Krashen, S. (1985). *The input hypothesis: issues and implications.* New York: Longman.

Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS.* New York: Routledge.

Leow, R. P. (2015). Conclusion: The changing L2 classroom, and where do we go from here? In Leow, R. P. (Ed.), *Explicit learning in the L2 classroom: A student-centered approach* (pp. 270–278). New York: Routledge.

Levy, M. & Hubbard, P. (2005). Why call CALL "CALL"? *Computer Assisted Language Learning, 18*(3), 143–149.

Liakin, D., Cardoso, W., & Liakina, N. (2015). Learning L2 pronunciation with a mobile speech recognizer: French /y/. *CALICO Journal, 32*(1), 1–25.

Liakin, D., Cardoso, W., & Liakina, N. (2017). The pedagogical use of mobile speech synthesis (TTS): Focus on French liaison. *Computer Assisted Language Learning, 30*(3–4), 348–365.

Lightbown, P. M. (2003). SLA research in the classroom/SLA research for the classroom. *The Language Learning Journal, 28*(1), 4–13.

Lin, C. H., Fisher, B. E., Winstein, C. J., Wu, A. D., & Gordon, J. (2008). Contextual interference effect: Elaborative processing or forgetting-reconstruction? A post hoc analysis of transcranial magnetic stimulation-induced effects on motor learning. *Journal of Motor Behavior, 40*(6), 578–586.

Lu, M. (2008). Effectiveness of vocabulary learning via mobile phone. *Journal of Computer Assisted Learning, 24*(6), 515–525.

Munro, M. J. & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *The Language Learning Journal, 45*(1)*,* 73–97.

Nation, P. & Wang Ming-Tzu, K. (1999). Graded readers and vocabulary. *Reading in a Foreign Language, 12*(2), 355–380.

Neri, A, Cucchiarini, C. & Strik, W. (2003). Automatic speech recognition for second language learning: How and why is actually works. In M. J. Solé, D. Recasens & J. Romero (Eds.), *Proceedings of the 15th international Conference on Phonetic Sciences, Barcelona, Spain* (pp. 1157–1160). Adelaide, AU: Causal Productions Pty Ltd.

Neri, A., Cucchiarini, C., Strik, H., & Boves, L. (2002). The pedagogy-technology interface in computer assisted pronunciation training. *Computer Assisted Language Learning, 15*(5), 441–467

Nusbaum, H. C., Francis, A. L., & Henly, A. S. (1995). Measuring the naturalness of synthetic speech. *International Journal of Speech Technology, 2*(1), 7–19.

Nye, P. W., Ingemann, F., & Donald, L. (1975). Synthetic speech comprehension: A comparison of listener performances with and preferences among different speech forms. *Haskins Laboratories: Status report on speech perception SR-41,* 117–126.

Ortega, L. (2013). *Understanding second language acquisition.* Abingdon, UK: Routledge.

Proctor, C. P., Dalton, B., & Grisham, D. (2007). Scaffolding English language learners and struggling readers in a universal literacy environment with embedded strategy instruction and vocabulary support. *Journal of Literacy Research, 39*(1), 71–9.

Rose, D. & Dalton, B. (2002). Using technology to individualize reading instruction. In C. C. Block, L. B. Gambrell, & M. Pressley (Eds.), *Improving comprehension instruction: Rethinking research, theory, and classroom practice* (pp. 257–274). San Francisco: Jossey-Bass.

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11,* 129–158.

Smith, B. (2004). Computer-mediated negotiated interaction and lexical acquisition. *Studies in Second Language Acquisition, 26,* 365–398.

Soler-Urzua, F. (2011). *The acquisition of English /ɪ/ by Spanish speakers via text-to-speech synthesizers: A quasi-experimental study* (Master's Thesis). Concordia University, Montreal, CA.

Stern, S. E., Mullennix, J. W., & Yaroslavsky, I. (2006). Persuasion and social perception of human vs. synthetic voice across person as source and computer as source conditions. *International Journal of Human-Computer Studies, 64*(1), 43–52.

Stevens, C., Lees, N., Vonwiller, J., & Burnham, D. (2005). On-line experimental methods to evaluate text-to-speech (TTS) synthesis: Effects of voice gender and signal quality on intelligibility, naturalness and preference. *Computer Speech & Language, 19*(2), 129–146.

Sundberg, R. & Cardoso, W. (2016). Aligning out-of-class material with curriculum: Tagging grammar in a mobile music application. In S. Papadima-Sophocleous, L. Bradley & S. Thouësny (Eds), *CALL communities and culture – short papers from EUROCALL 2016, Limassol, Cyprus* (pp. 440-444). Dublin, IE: Research-publishing.net.

Tanaka, T. (2009). Communicative language teaching and its cultural appropriateness in Japan. *Doshisha Studies in English, 84*, 107–123

Thomson, R. I. (2011). Computer assisted pronunciation Training: Targeting second language vowels perception improves pronunciation, *CALICO Journal, 28*(3), 744–65.

Thomson, R. I. (2012). Improving L2 listeners' perception of English vowels: A computer–mediated approach. *Language Learning, 62*(4), 1231–1258.

Trofimovich, P., Collins, L., Cardoso, W., White, J., & Horst, M. (2012). A frequency-based approach to L2 phonological learning: Teacher input and student output in an intensive ESL context. *TESOL Quarterly, 46*(1), 176–186.

VanPatten, B. (2007). Input processing in adult second language acquisition. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition* (pp. 115–135). New Jersey: Lawrence Erlbaum

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics, 28*(1), 46–65.

Appendices

Appendix A: Self-assessment form

LANGUAGE BACKGROUND

- **Place of birth:** _____
- **Native Language:** _____
- **Level at the school:** _____

- Do you have or have you had any **hearing problems**? **YES** **NO**

- What other languages do you know?

Language	**Proficiency**
1.	Beginner Intermediate Advanced Native
2.	Beginner Intermediate Advanced Native

ENGLISH PROFICIENCY AND EXPERIENCE

How do YOU evaluate your overall **proficiency in English**: Beginner Intermediate Advanced

Approximately what percent of the time do you **SPEAK English** in your daily life?

0% 10 20 30 40 50 60 70 80 90 100%

Approximately what percent of the time do you **LISTEN to English** (radio, internet, TV, etc.)?

0% 10 20 30 40 50 60 70 80 90 100%

On a scale of 1-10, how would you **rate your LISTENING ability in English**?

0 1 2 3 4 5 6 7 8 9 10

On a scale of 1-10, how would you **rate your SPEAKING ability in En**glish?

0 1 2 3 4 5 6 7 8 9 10

Approximately what percent of the time do you **interact with other native English speakers**?

0% 10 20 30 40 50 60 70 80 90 100%

Appendix B: Dictation Answer Sheet

*See Appendix G for the transcripts

In this task, you will listen to 10 sentences and then write what you heard.

Sentence 1: _____

_____

Sentence 2: _____

_____

Sentence 3: _____

_____

Sentence 4: _____

_____

Sentence 5: _____

_____

Sentence 6: _____

_____

Sentence 7: _____

_____

Sentence 8: _____

_____

Sentence 9: _____

_____

Sentence 10: _____

_____

Appendix C: Comprehension Tests

*See Appendix F for the transcripts

Short Story #1 COMPREHENSION QUESTIONS

Based on the story you heard, choose the BEST answer for the following questions.

1) How many people were on the plane?
a) Four
b) Five
c) Only the pilot
d) The plane was empty

2) What started to happen to the plane?
a) The plane lost one engine
b) A fire started
c) The plane started to shake
d) There were snakes on the plane

3) Who left the plane first?
a) The nun
b) The president
c) The schoolboy
d) The pilot

4) What was the nun holding when the plane started to shake?
a) A newspaper
b) The Bible
c) A cross
d) A cat

5) If there are only 4 parachutes for 5 people, how did both the nun and schoolboy both have one?
a) The pilot miscounted the parachutes
b) The nun prayed for a miracle and an extra parachute appeared
c) Someone jumped without a parachute
d) The schoolboy found an extra parachute

6) What probably happened after?
a) The nun and the schoolboy survived, but the professor died
b) The professor survived, but the nun and the schoolboy died
c) Everybody died
d) Everybody survived

Short story #2 COMPREHENSION QUESTIONS

Based on the story you heard, choose the BEST answer for the following questions.

1) Why did the woman go to the store?
a) Because she worked at the store
b) Because someone asked her to come inside
c) Because she wanted to buy a present for her mother
d) Because her mother was in the pet shop

2) What color was the bird?
a) Red and blue
b) Red and black
c) Green and blue
d) Green and black

3) Why was the bird so special, according to the woman?
a) Because it had feathers of gold
b) Because it could read the future
c) Because it could talk and sing in different languages
d) Because it was a special dish

4) How much did the bird cost?
a) It cost $15,000 dollars
b) It cost $1,500 dollars
c) It cost $15 dollars
d) It cost $50,000 dollars

5) What did her mother do with the bird?
a) She built a beautiful cage for the bird
b) She ate the bird
c) She taught the bird some songs
d) She returned the bird to the pet shop

6) What do you think probably happened after?
a) The daughter bought another bird for her mother
b) The daughter was very angry
c) The mother went to the pet shop to buy more birds
d) The pet shop wanted the bird back

Appendix D: 6-point Likert Scale for Comprehensibility, Naturalness, and Pronunciation Accuracy Ratings

*See Appendix H for the transcripts

RATING TABLE						
How EASY was the voice to UNDERSTAND?						
Very Hard						Very Easy
0	1	2	3	4	5	6
How NATURAL was the voice?						
Very Unnatural						Very Natural
0	1	2	3	4	5	6
How CORRECT was the pronunciation?						
Very Poor/Incorrect						Very Good/Correct
0	1	2	3	4	5	6

Appendix E: Aural Identification Answer Sheet

*See Appendix I for the transcripts

In this last task, you will listen to 16 sentences. This time, however, instead of being asked to rate them, you will be asked if you heard a certain sound in them. The sound target you will be focusing on is the **past tense** *-ed.* This sound can take one of three forms:

1. /t/ as in *walked*
2. /d/ as in *played*
3. /ed/ as in *waited*

When listening to these sentences, please listen carefully and mark either PAST or NOT PAST.

Practice. Please circle whether you heard the past tense *-ed* sound or not.

Practice Sentence 1:

PAST NOT PAST

Practice Sentence 2: PAST NOT PAST

Let's start: Circle whether you heard the past tense *-ed* sound or not

Sentence 1: PAST NOT PAST	Sentence 9: PAST NOT PAST
Sentence 2: PAST NOT PAST	Sentence 10: PAST NOT PAST
Sentence 3: PAST NOT PAST	Sentence 11: PAST NOT PAST
Sentence 4: PAST NOT PAST	Sentence 12: PAST NOT PAST
Sentence 5: PAST NOT PAST	Sentence 13: PAST NOT PAST
Sentence 6: PAST NOT PAST	Sentence 14: PAST NOT PAST
Sentence 7: PAST NOT PAST	Sentence 15: PAST NOT PAST
Sentence 8: PAST NOT PAST	Sentence 16: PAST NOT PAST

Appendix F: Short Stories' Transcripts

Airplane

A pilot and four passengers were flying in an airplane. The passengers were the President of the United States, a university professor, a schoolboy, and a nun. All of a sudden, the plane started to shake. The passengers looked at each other nervously. The pilot shouted: "Passengers! Your attention, please. We are going down! I counted the parachutes and I am sorry, but there are only four for the five of us." Then the pilot took a parachute, jumped out and landed safely.

Now there were only three parachutes.

"I am the most important man in the world," said the President and he took a parachute. "I must live! I must live!" he repeated. He then jumped out of the airplane. He, too, landed safely.

Now there were only two parachutes.

"I am the most intelligent man in the world," the university professor stated. "I, too, must live." He took a parachute and then jumped out of the plane too.

The nun folded the newspaper she was reading and said to the schoolboy, "You take the last parachute, my son. I am ready to die." She smiled, thinking of her new life in heaven.

"It's OK," the schoolboy answered. "There are two parachutes left."

"How can that be?" the surprised nun demanded. "There were only four parachutes for the five of us."

"That's right," said the schoolboy, "but the most intelligent man in the world jumped out of the airplane with my backpack."

Happy Birthday

A rich woman wanted to send her mother a very nice birthday present. One day, she walked past a pet shop in New York City. She saw a beautiful red and blue bird in the window. She tapped the window and smiled at the bird. She hoped that the bird was for sale. Opening the door, she went inside. The bird began to sing when the woman stopped next to the cage. She listened to the bird's song. It was beautiful! It could talk too, and it sang songs in Portuguese and English. She thought that the bird was very sweet and intelligent.

The woman decided that she wanted to buy the bird for her mother, but she had a couple of questions about the bird. She saw an employee and asked for help. The employee was very friendly. He answered many questions about the marvelous bird and the woman decided to buy it. It cost fifty thousand dollars! She opened the zipper on her purse and took out a credit card.

The next morning, the store delivered the bird to the woman's mother. That afternoon, the rich woman phoned to talk to her mother, "Mama," she said, "do you like the bird?"

"I'm eating it right now," her mother said. "It is delicious! Thank you so much."

Appendix G: Dictation/Intelligibility Task – Transcript of Target Sentences

1. A four-year-old boy sat in the doctor's waiting room with his mother.
2. He saw a pregnant woman on the other side of the room.
3. Is the baby in your stomach?
4. If he is such a good baby, then why did you eat him?
5. Last Christmas, Jimmy received the best present: it was a parrot.
6. Jimmy heard the parrot say some very bad words.
7. Jimmy was so frustrated that he decided to punish the bird.
8. He carried his parrot into the kitchen and put it in the freezer.
9. He did not know why the parrot stopped saying bad words after only a few minutes in the freezer.
10. May I ask what the chicken did wrong?

Appendix H: Rating Task Sentences' Transcript

1. He placed the glasses on his nose and looked up.
2. When he arrived, he saw that the front door was open.
3. She quickly opened the box and found the pictures and the letter.
4. I looked for your picture, but I can't remember which girl you are.
5. He stood up and walked to the chair where she was sitting.
6. The boy watched the clock ticking on the wall.
7. He talked to his mother very politely and said very nice things.
8. His mother and father explained that bad words were not polite.
9. The boy stepped back from the fence and rolled up his pants.
10. The girl put her hand into her pocket and pulled out a handful of change.
11. The teacher talked for twenty minutes about school and being good students.
12. My teacher asked me to please sit down.

Appendix I: Aural Identification Task – Transcript of Target Sentences

Practice 1: I ordered a large pizza.

Practice 2: I water my garden.

1. I called my mother.
2. I visit my cousin Sam.
3. I talked with Jeff in the hallway.
4. I grilled the hamburgers.
5. I corrected my math homework.
6. I jumped in the freezing lake in winter.
7. I study English for 4 hours.
8. I invited him to dinner.
9. I finish my homework at 9pm.
10. I receive many presents on my birthday.
11. I opened the door for her.
12. I fixed the problems around the house.
13. I hated the movie.
14. I danced to the music.
15. I waited two hours for my friend.
16. I painted some pictures.

Appendix J: Interview Sample Questions

*Questions asked in the participants' native language, Portuguese.

1. Did you notice anything different between the voices you heard in the study?

2. What did you think about the computer-generated speech that you heard? Was it easy to understand?

3. How good do you think the computer-generated speech would be as a learning tool? For pronunciation?