A Comparison of Synthetic and Human Speech: an Evaluation by English as a Foreign Language Students in a Public Costa Rican University

Recibido: 12 de febrero, 2023
Aceptado: 13 de noviembre, 2023
Por: William Charpentier-Jiménez[1], Universidad de Costa Rica,
ORCID: https://orcid.org/0000-0002-8554-7819

Abstract

The possible role of text-to-speech (TTS) audio for pedagogical purposes has not been fully explored. This study examines ESL students' perceptions of artificial intelligence and human voices. It also explores students' opinions on listening instruction. The investigation was conducted from April to September 2022 and involved 36 TESOL students enrolled in a BA in English or English teaching at a Costa Rican public university. It used a quantitative survey design. The researcher gathered student responses through a survey designed to collect students' perceptions of computer-generated voices, human voices, and listening instruction. The data were quantitatively analyzed using descriptive statistics. Data analyses indicate that: 1) students find human voices more appealing than artificial intelligence voices; 2) students find female voices more appealing than male voices when a computer generates them; 3) artificial intelligence voices share some characteristics that students find more appealing; and 4) current listening instruction policies and materials should be reexamined in the language program. Consistent with the reviewed literature, these findings demonstrate that although TTS does not appeal to students as much as human voices, a part of the population finds computer-generated voices appealing. The analysis also suggests that some students cannot fully discern between computer-generated and human voices; thus, their use may be appropriate in some contexts. Finally, these findings confirm that listening instruction policies and materials should be revised to improve students' language acquisition processes.

William Charpentier-Jiménez. A Comparison of Synthetic and Human Speech: an Evaluation by English as a Foreign Language Students in a Public Costa Rican University. *Revista Comunicación*. Año 44, volumen 32, número 2, junio-diciembre, 2023. Instituto Tecnológico de Costa Rica. ISSN: 0379-3974/e-ISSN1659-3820

Comparación del Habla Sintética y Humana: una Evaluación de Estudiantes de Inglés como Lenguas Extranjera en una Universidad Pública Costarricense

Resumen

El posible papel de audios texto-a-voz (TTS) para usos pedagógicos no ha sido completamente explorado. Este estudio examina las percepciones de estudiantes

PALABRAS CLAVE:
Inteligencia artificial, enseñanza de una lengua extranjera, educación superior, material pedagógico de escucha, texto-a-voz

KEY WORDS:
Artificial intelligence, Foreign language instruction, Higher education, Teaching listening materials, Text-to-speech

1 William Charpentier-Jiménez ha trabajado para la Universidad de Costa Rica por más de catorce años como profesor y coordinador de varias secciones y proyectos. Ha impartido clases tanto en el Bachillerato en la Enseñanza del Inglés así como en el Bachillerato en Inglés, la Licenciatura en la Enseñanza del Inglés y la Maestría en la Enseñanza del Inglés. Posee una Maestría en Lingüística Aplicada y una Maestría en Administración Universitaria, ambas de la Universidad de Costa Rica. Sus principales intereses de investigación se relación con la adquisición de vocabulario, aprendizaje de la lengua asistido por computadora y dispositivos móviles, y la pronunciación. Contacto: william.charpentier@ucr.ac.cr

de ILE acerca de las voces humanas y de inteligencia artificial. Asimismo, explora las opiniones de estudiantes sobre la instrucción de la escucha. Esta investigación se llevó a cabo de abril a setiembre de 2022 e incluyó a 36 estudiantes de ILE matriculados en un Bachillerato en Inglés o Enseñanza del Inglés en una universidad pública costarricense. Se utilizó un modelo cuantitativo de encuestas. El investigador recolectó las respuestas mediante una encuesta diseñada para recabar las percepciones del estudiantado acerca de las voces generadas por computadora, las voces humanas, y la instrucción de la escucha. Los datos fueron analizados de manera cuantitativa utilizando estadística descriptiva. El análisis de los datos indica que: 1) el estudiantado encuentra las voces humanas más atractivas que las voces generadas con inteligencia artificial; 2) el estudiantado considera las voces femeninas más atractivas que las masculinas cuando son generadas por computadora; 3) las voces generadas por inteligencia artificial comparten algunas características que el estudiantado encuentra más atractivas; y 4) las presentes políticas y materiales para la instrucción de la escucha deben ser reexaminadas en el programa de idiomas. Consistente con la literatura revisada, estos resultados demuestran que aunque las voces TTS no llaman tanto la atención del estudiantado como las voces humanas, una parte de la población considera las voces generadas por computadora interesantes. El análisis también sugiere que una parte del estudiantado no puede discernir en su totalidad entre voces humanas y generadas por computadora; por lo tanto, su uso puede ser apropiado en algunos contextos. Finalmente, los resultados confirman que las políticas y los materiales para la enseñanza de la escucha deben ser revisados para mejorar los procesos de adquisición del lenguaje del estudiantado.

INTRODUCTION

In recent years, humanity has moved beyond viewing technology as a tool to recognizing its potential as a creative entity capable of imitating human characteristics. With the advent of artificial intelligence (AI), self-driving cars, automated assistants, computer-generated text, and synthetic speech have become commonplace in our daily lives (Adamopoulou & Moussiades, 2020; Luo et al., 2022). However, the potential of these inventions in educational settings has yet to be fully explored. As AI becomes more capable of processing language and closely interacting with humans, its application in language learning settings stands out. For example, TTS software could help the visually impaired, produce voices with realistic accents, or provide fully customized, appropriate audio input for language learners. However, many professors remain unaware or passive about the possible advantages or disadvantages of this technology. Consequently, this hinders students from accessing a potential source of input tailored to their specific requirements, interests, and language level.

Although this paper focuses on the English language, public institutions in Costa Rica frequently offer other languages as a major (e.g., French) or as required or elective courses for other majors. Therefore, the potential application of TTS may impact several language programs. Regardless of the potential benefits of implementing TTS in ESL settings (Bione et al., 2017; Cardoso et al., 2015; Craig & Schroeder, 2019; Hillaire et al., 2019; Kang & chatGPT, 2023), currently, the only discussion about TTS systems in the English major has centered on using them to read abstracts. No other language program in the university where this study took place has considered using TTS synthesis, and no attention has been paid to the use of computer-generated voices in educational settings, especially in ESL settings where listening materials are abundant but sometimes difficult to find or do not fully adapt to the course requirements.

This study aims to identify students' perceptions of synthetic and human voices. The findings may explain the possible advantages of incorporating TTS voice recordings in the language class. By comparing human voices to synthetic voices, the findings will also aid in understanding students' preferences and the differences and similarities between both. Thus, the results may encourage the development of audio materials, audiobooks, audio instructions, and auditory aids for visually impaired students.

The findings of this research will potentially benefit two populations. On the one hand, professors will have research-based evidence to know whether or not using TTS is appropriate. They can also consider using TTS in some circumstances but not in others. For example, professors may determine that TTS is useful for assisting students with visual impairments or creating audio instructions for listening tasks but not good enough for longer text passages. Also, professors may incorporate audio clips to create other kinds of materials (for example, audio prompts or word lists in audio) or tasks where students respond to questions or react to a short listening passage. On the other hand, students may benefit

from being exposed to materials that are more contextualized to their needs, more adequate to their level, and more appropriate to their interests. Students can also be exposed to AI speakers of various accents, ages, or genders if needed. Therefore, having this variety and adaptation may enrich the class dynamics in the ESL class.

This paper is divided into five distinct sections. The introduction describes the importance and potential benefits of including TTS audios in the context of English as a second language or foreign language (ESL/EFL). The literature review presents the most relevant concepts of artificial intelligence (AI) and natural language processing (NLP). It also introduces some core concepts related to the human voice, TTS theory, and listening instruction. The methods section describes the participants, materials, methodology, procedure, and data collection and interpretation steps. The results section offers a statistical analysis of the data collected. Finally, the discussion summarizes some possible limitations and proposes the main results of the study and their implications in the field of language teaching, particularly listening instruction.

Aims

The article aims to compare students' perceptions of synthetic and human speech. This article also focuses on the perceived differences or similarities between male and female voices.

REVIEW OF THE LITERATURE

The role of listening instruction has been extensively studied in recent years. However, to the best of the researcher's knowledge, studies comparing AI-generated audio and human-created audio in ESL settings are scarce. This literature review summarizes some of the main concepts related to this study. It does not intend to be comprehensive but to provide an overview of the central aspects of speech synthesis and language instruction, particularly the listening skill.

Artificial Intelligence and Natural Language Processing

Artificial intelligence is the simulation of human intelligence in machines designed to think and act like humans. It also involves creating machines that can learn from data, make predictions, make decisions, and perform tasks that would typically require human intelligence, such as visual perception, speech recognition, decision-making, and translating (Abbott, 2020; Arora, 2022; Cameron, 2019; Jeste et al., 2020; Kent, 2022). There are various types of AI. Narrow or weak AI is designed to perform a single task (Gulson et al., 2022; Kindersley, 2023), while general or strong AI can perform any intellectual task that a human can (Kindersley, 2023; Mitchell, 2019). Artificial intelligence is used in various applications, such as self-driving cars, virtual personal assistants, and biometric authentication methods. In addition, AI research aims to create systems capable of performing tasks that typically require human intelligence, such as understanding natural language, recognizing images, playing games, and solving complex problems (Jeste et al., 2020).

Natural Language Processing (NLP) is a subset of artificial intelligence that focuses on the interaction of computers and humans through natural language (Kochmar, 2022; McRoy, 2021). It involves the development of models and algorithms that can analyze, comprehend, and produce human language. Natural Language Processing is used in various applications, such as sentiment analysis, online searches, predictive text, and machine translation (Adamopoulou & Moussiades, 2020; Luo et al., 2022). It has also been instrumental in advancing virtual assistants and chatbots. In addition, NLP techniques are based on a combination of computer science, linguistic theory, and machine learning (Raaijmakers, 2022). The goal of NLP is to create systems that can accurately compute and analyze large amounts of data and use this information to perform specific tasks. Overall, NLP is a rapidly growing field that has the potential to revolutionize how computers and humans interact and has a wide range of practical applications, including customer service, marketing, healthcare, etc.

The Human Voice

The human voice is the sound produced by the vibration of the vocal folds in the larynx. Sound waves are produced by the vibration of the vocal folds, which travel through the oral and nasal cavities to produce speech or singing; these waves interact with our articulators (tongue, jaw, teeth, etc.) to produce specific sounds

(Calais-Germain & Germain, 2016). The human voice is a powerful and unique tool for communication and self-expression. It can convey a wide range of emotions and has been used throughout history for interacting, storytelling, singing, and other forms of artistic expression (Karpf, 2006).

The study of the human voice, including its production and perception, is known as voice science or phonetics (Akmajian et al., 2017). Voice science deals with the sound and quality of the voice, which in turn is influenced by several factors, including age, gender, physical attributes, and emotional state. This field of study is essential for understanding how the voice works and developing techniques for improving vocal health and performance, helping people with trouble speaking, or developing techniques and strategies to help students learn a new language.

In addition to its role in communication and expression, the human voice also plays an essential role in identity and socialization, as it is frequently used to convey personal and cultural information (Norton & Toohey, 2011). Thus, the human voice is a complex and unique aspect of human physiology and behavior and continues to be studied by scientists and artists alike. It provides essential information such as gender, personality, accent, race, and emotion, among other aspects (Nass & Brave, 2005). However, the role of the voice as an instrument that carries a message has been frequently overlooked (Karpf, 2006). Listening exercises frequently focus more on the quality of the audio in general than on the characteristics of the voice, and research about the role of the human voice in learning remains scarce (Craig & Schroeder, 2019).

Assistive and text-to-speech technology

Assistive technology refers to tools, devices, or software created to help people with disabilities perform tasks that they would otherwise be unable to perform or may complete with difficulty (Emiliani & Association for the Advancement of Assistive Technology in Europe, 2009). These technologies can aid people with several disabilities, including physical, sensory, and cognitive impairments (Bouck, 2017; Cook, 2019; Dell et al., 2017; Green, 2018). Examples of hardware assistive technology include adaptive computer hardware, such as large print keyboards, mouse devices, screen magnifiers, and adapted joysticks for individuals with mobility or dexterity issues. Examples of software assistive technology include screen readers and TTS software for individuals who are blind or have low vision.

Although screen readers and TTS systems are similar, they also have some differences. A screen reader is a type of assistive technology that reads out loud the text on a computer screen. It is primarily designed to help individuals who are blind or have low vision access the information and functions of a computer (Evans & Blenkhorn, 2008). In this case, the program reads what is already on the screen. Text-to-speech is an advanced technology that converts written text into speech. One of its main goals is to be very similar or even indistinguishable from the human voice (Dutoit, 1997). It uses natural language processing and speech synthesis to generate human-like speech from input text (Taylor, 2009). The output speech can be played back using speakers or headphones or stored as an audio file. For instance, unlike screen readers, a user can deliberately enter text to be read. This first user can modify the text and how it will be presented to the end user. Text-to-speech technology is commonly used for accessibility purposes, for individuals with visual impairments, and for various applications in fields such as education, entertainment, and business (Narayanan & Alwan, 2005).

Text-to-speech technology breaks down written text into words and phrases and then uses a computer-generated voice to read them aloud. The process of TTS typically involves the following stages: text analysis (the text is analyzed and processed to determine pronunciation, rhythm, and stress patterns), voice synthesis (a computer-generated voice is created by concatenating or piecing together segments of pre-recorded speech), and speech production (the processed text is combined with the generated voice to produce spoken language) (Hersh et al., 2008; Holmes & Holmes, 2001).

Several studies have found potential benefits from using TTS in language classes or have found no significant differences between using a synthetic or human voice (Bione et al., 2017; Cardoso et al., 2015; Craig & Schroeder, 2019; Hillaire et al., 2019; Kang et al., 2008) and the possibility of more interactive models where people can keep conversations in real time with machines (Ku-

mar et al., 2023). In addition, TTS systems may use various techniques to improve the quality and naturalness of the generated speech, such as adjusting the rhythm and intonation to match that of a human speaker or adding natural-sounding pauses and inflections. Since this technology is rather new, it is constantly evolving and improving (Chen et al, 2023; Wang et al., 2023). Therefore, the accuracy and quality of TTS systems can vary widely, depending on factors such as the complexity of the text, the quality of the voice synthesis, and the sophistication of the TTS algorithms used.

Teaching Listening

Teaching listening skills involves providing students with opportunities to practice and develop their ability to understand spoken language. It also includes strategies such as providing opportunities for authentic listening, using varied listening materials, and incorporating interactive activities (Brace et al., 2006; Ur, 2012). Teaching listening skills also requires dedication and a focus on the process and the outcome. Thus, regular practice and ongoing feedback are essential for helping students improve their listening abilities.

Listening has historically been viewed as a receptive skill (Field, 2011; Harmer, 2013). To understand a listening passage, the listener uses their linguistic abilities and schemata. In this regard, one of the main difficulties for ESL students is making sense of the sound system of English, especially if they are learning it as adults (Field, 2011; Nation & Newton, 2009). On the other hand, the topics used in language classes should consider students' schemata. A schema is a cognitive framework or mental model that helps us organize and understand information (Brown & Lee, 2015; Harmer, 2013). Schemata can refer to general concepts or mental structures about the world or specific knowledge structures about a particular topic or situation. For example, we may have a schema about what a typical car looks like, which helps us understand and categorize new information about cars we encounter in real-life situations or through written, pictorial, or audio messages. Therefore, the learner's linguistic proficiency and schemata are crucial when decoding the message.

In addition, some other aspects may constrain students' listening comprehension. Some of these limitations are related to language features or contextual characteristics of the message. For example, the speed of delivery (Brown & Lee, 2015; Ur, 2012) or the speakers' accent, especially if no adequate training has been previously provided (Charpentier-Jiménez, 2019; Derwing & Munro, 2015; Field, 2011; Harmer, 2007), may limit students' processing time and frustrate their attempts to decode the message. Additionally, the type of vocabulary (Hadfield & Hadfield, 2008; Watkins, 2010) and the level of formality (Hadfield & Hadfield, 2008) could slow down students' ability to comprehend the message. On the one hand, the words, expressions, or grammar used could be too specific, elaborate, or technical for students to understand. On the other hand, language could be too colloquial and culturally bound, making understanding the message more challenging.

Another aspect to consider is the message and its characteristics. For example, audio input should present students with authentic input while considering various task types and audio formats (Brown & Lee, 2015; Burgess & Head, 2005; Celce-Murcia et al., 2010). Content is another aspect professors should examine (Harmer, 2007). These aspects make finding voice recordings more difficult for professors. Despite the myriad possibilities the Internet brings, audio recordings do not always adapt to students' levels, the desired task, the appropriate accent, or the content under study. Moreover, professors should consider aspects like length, audio recording quality, or any other aspect that interferes with the message, such as background noise (Watkins, 2010), since the audio input should provide students with an appropriate model to imitate (Patel & Jain, 2008).

Finally, the advancement of AI and text-to-speech systems have proven effective in improving language learning. A study by Al-Jarf (2022) highlighted notable improvements in decoding skills, reading fluency, and pronunciation accuracy when using these tools, although there was no significant enhancement in vocabulary knowledge. Additionally, the integration of AI-driven techniques in ELT has been instrumental in bossing motivation and fostering heightened learner engagement. As highlighted by Anis (2023), learners experience heightened involvement due to the effects of adaptive instruction, intelligent tutoring systems, and personalized learning applications. These innovative

approaches not only stimulate motivation but also encourage active participation in language-related activities. Furthermore, as Moybeka et al. (2023) emphasize, text-to-speech applications serve as pivotal tools in dismantling language barriers, leading to a more inclusive and equitable approach to English language education. TTS also offers a unique advantage, assisting students in refining their listening and reading proficiencies (Hartono et al., 2023). Text-to-Speech tools could be a valuable asset in acquainting students with a diverse range of accents, further enriching their auditory experience and understanding of the language (Fitria, 2023).

This literature review presents some main concepts related to using TTS in ESL classes. The researcher must grant that some concepts related to TTS systems or listening instruction have been purposely left aside as they do not directly relate to the objective of this study. However, this omission does not limit or impair the findings of this paper.

METHODS

Participants

This study includes Costa Rican university students enrolled in their second language course. The researcher visited the students' oral classes to invite them to participate. The participants were selected because they were currently enrolled in an oral course in their second academic year. Their proficiency level corresponds to B1-B2. Thirty-six participants were willing to participate; however, they did not receive monetary compensation for their participation. All participants speak Spanish as their first language.

Materials

The materials include written consent, a listening script (see Appendix 1), four different audios (see Appendix 3), the software, the necessary equipment for the listening part, and an electronic survey (see Appendix 2) to collect participants' answers. The written consent was sent to participants electronically before their participation, and a checkbox labeled "agree to terms and conditions" was included to certify voluntary participation in the study. The listening script used was *Comma Gets a Cure*, a diagnostic passage for dialect and accent that can be used freely without special permission. At no point in the study did students have access to the script. The four different audios included this same passage. To record the audio, the software Speechelo was used. Speechelo is an AI-enabled TTS and voiceover, paid software that turns text into human-sounding voiceovers (BlasterOnline, 2023). It can also create audio in 23 languages. It was chosen because of its quality and the number of audios it has available. Two of the audios were read by a male and female human, both native American English speakers. Students were not informed that some audios could be computerized as this could have biased their perception. The other two audios were read in American English by a male and female AI voice using Speechelo. All audios were encoded in an MP3 format. Participants listened to the audio using noise-canceling, over-the-ear headphones, the Bose QC35 Series II, which guarantees optimal listening conditions. These headphones were wirelessly connected to a different audio system to avoid any interference with students' answers. Finally, the survey was divided into four sections: a) demographic information, b) participants' perceptions of voice recordings in English classes, c) the evaluation of the AI or human audio, and d) an optional open-ended question. The survey used two question formats: forced-choice and open-ended questions. Except for the open-ended question, items included Likert scales for all sections. For example, some items asked the participants to rate the audio quality in their English classes. These items were placed on a 5-point Likert scale that ranged from 1 (Very poor) to 5 (Very good). This format, or a similar one, was also used for other questions.

The last part of the survey contained one optional, open-ended question. This question invited participants to add any other comments they believed were relevant to the study. The total time to complete the survey was estimated at 10 to 15 minutes.

PROCEDURE

This study used a quantitative survey design. First, the researcher selected an appropriate text to create the audio recordings. The text was selected because it is copyright-free and normally used in language analysis. The researcher then pilot-tested several female and male AI voices with ten participants from the same affiliation

as the target population. This stage aimed to extract the two voices that sounded more human-like. The AI voices chosen (Mathew and Grace) were fed the proposed text. These two voices were chosen from a list of 17 voices offered by the software. Although Speechelo allows the user to add breathing and pauses, among other changes, the audios were not modified in any way. The human voices were professional voiceover actors. The speakers also read the same text, and their voices were in no way altered.

After preparing the materials, the researcher created the survey. The survey included sections about participants' demographic information, their perception of audio quality in English classes, a list of ten descriptors to evaluate the four voiceovers, and an open-ended question. To explore participants' perceptions of audio recordings, the list of ten descriptors was extracted from a list of 17. This list was compiled by the researcher, considering the most common characteristics associated with vocal features (Memon, 2020, Paz et al, 2022). Some items from the initial list were discarded since they did not fit the study's scope (i.e., background noise, length, and volume, among others). By default, some of these features were either objectively the same in all audios or could be adjusted by the participants. All students had access to a sample survey before their appointment.

Finally, during the data-gathering stage, participants were summoned to a vacant office with a silent environment. Students were able to choose their appointments at their own convenience. The researcher provided written and oral instructions to all participants. Participants used noise-canceling, over-the-ear headphones to minimize any background noise during this stage. Although participants could listen to the audio more than once, no student asked to listen again. Participants' answers were collected through an anonymous electronic survey that was partially completed while listening to the audio. Other sections did not require the audio to be completed.

Data Processing and Analysis

The original data set in Excel format (xls) was subjected to computational analysis using the statistical package for social sciences (SPSS) Version 26. The data was derived from participants' survey answers. The analysis included descriptive statistics, where percentages, nominal data, and the standard deviation, among other basic statistics, were performed to compare participants' opinions about the audios and their listening training.

ANALYSIS OF THE RESULTS

The following summary of the results presents the main findings of the study in four distinct sections. The first section includes the participants' demographic information. The second section compares the four voiceovers based on participants' ratings. The third section summarizes the main features under analysis and their ratings. Finally, the fourth section describes participants' general perceptions of the audios used during English classes and the type of listening instruction they received.

Demographic Information

Of the 36 study participants, 27 (75%) were females and 8 were males (22.22%). One participant (2.78%) chose to be identified as non-binary. Overall, 33 participants (91.67%) reported being between the ages of 18 and 24, while two (5.56%) were between 25 and 34. Only one participant (2.78%) was between 35 and 44. All participants are native Spanish speakers and study English as a foreign language. Regarding studies, the study participants are enrolled in the BA in English (n = 29, 80.56%) or English teaching (n = 6, 16.67%). Only one participant reported studying both majors (n = 1, 2.78%). The two majors share the same core language courses, including oral courses. All of the participants are currently in their second or third year.

Participants' ratings of voiceovers

The following analysis delves into the realm of participant voiceover preferences. Table 1 shows that participants' voiceover ratings can be analyzed from two perspectives. First, participants had a slight preference for female voices. Although the difference was almost non-existent when comparing human voices, the female AI was more than six points above the male AI. Second, participants showed a marked preference for human voices. Even though all maximum grades were at or above 90, the minimums for AI voices were below 55, while human voices exceeded the 70 threshold. The

standard deviation also shows that, when evaluating human voices, ratings tend to be more uniform. However, ratings are more spread when evaluating AI, indicating that, although AI voices ranked lower than human voices, they appealed to part of the population. This was especially evident when overlapping those results with the mean and maximum grades.

Table 1. ***Summary of participants' perceptions of each voiceover: mean and standard deviation***

	Min.	Max.	$\bar{X}$	Median	SD
Female AI	54.00	98.00	78.44	76.00	14.24
Male AI	54.00	90.00	72.22	74.00	12.35
Female Human	76.00	96.00	86.44	86.00	6.69
Male Human	74.00	98.00	86.22	90.00	8.74

Note. N = 36. Min. = Minimum; Max. = Maximum; $\bar{X}$ = arithmetic mean; SD = Standard Deviation

Source: Compiled by the author based on survey responses.

Participants' voiceover rating per criteria

To analyze participants' perceptions of each audio, ten criteria were chosen. As previously stated, some criteria from the initial list of 17 were discarded. While listening to each audio, participants used a five-point Likert scale to rate one of the audios that were evenly and randomly assigned to them. Table 2 presents a summary of the main findings of this section.

Table 2. ***Summary of participants' ratings of each criterion: means of raw data and percentage***

Criteria	Female AI	Male AI	Female Human	Male Human
Intonation (monotonous – varied)	26 (57.78%)	22 (48.89%)	36 (80.00%)	40 (88.89%)
Voice quality (unclear – clear)	45 (100%)	39 (86.67%)	43 (95.56%)	42 (93.33%)
Voice quality (harsh – pleasant)	37 (82.22%)	27 (60.00%)	38 (84.44%)	38 (84.44%)
Voice quality (lifeless – enthusiastic)	30 (66.67%)	32 (71.11%)	38 (84.44%)	36 (80.00%)
Speed (paused – fluent)	41 (91.11%)	43 (95.56%)	42 (93.33%)	41 (91.11%)
Speed (unvaried – varied)	29 (64.44%)	33 (73.33%)	36 (80.00%)	34 (75.56%)
Vocal variety (does not convey emotion-conveys emotion)	26 (57.78%)	27 (60.00%)	34 (75.56%)	32 (71.11%)
Vocal variety (unfriendly – friendly)	38 (84.44%)	37 (82.22%)	39 (86.67%)	40 (88.89%)
Vocal variety (strained – natural)	39 (86.67%)	30 (66.67%)	42 (93.33%)	41 (91.11%)
General audio quality (unintelligible – clear)	42 (93.33%)	35 (77.78%)	41 (91.11%)	44 (97.78%)

Note. N = 36.

Source: Compiled by the author based on survey responses.

According to Table 2, some criteria show more contrast, while others are more similar. In terms of similar characteristics, speed (paused – fluent) (SD = 2.13) and vocal variety (unfriendly – friendly) (SD = 2.87) are three or fewer points apart from each other. In both cases, participants perceive the friendliness and fluency of the voice as good and very good, respectively. On the other hand, some criteria were different. For example, according to the participants' answers, intonation (monotonous – varied) (SD = 18.68) and vocal variety (strained-natural) (SD = 12.17) are characteristics that show great variation, favoring human voices. Another characteristic worth mentioning is voice quality (harsh-pleasant) (SD =11.90). In this last case, the variation occurred mainly because of the perceived harshness of the male AI voice.

In addition, some criteria had higher or lower marks overall. For example, voice quality (unclear – clear) (93.33%) and speed (slow – fluent) (93.33%) were the two highest-ranked criteria for AI voices. In the case of human voices, voice quality (unclear – clear) (94.44%) and general audio quality (unintelligible – clear) (94.44%) were the highest. This shows that overall voice quality (unclear – clear) was the characteristic that appealed most to participants. On the other hand, AI voices ranked the lowest in intonation (monotonous-varied) (53.33%) and vocal variety (does not convey emotion – conveys emotion) (58.89%), while human voices ranked the lowest in speed (unvaried – varied) (77.78%) and vocal variety (does not convey emotion – conveys emotion) (73.33%). On average, vocal variety (does not convey emotion – conveys emotion) was the characteristic participants found more unappealing.

Participants' perceptions of audio use and instruction

Participants' perceptions of audio quality and instruction were analyzed based on six survey questions. First, the survey considered listening instruction. Eight participants (22.22%) ranked listening instruction as very good, while 17 (47.22%) ranked it as good. Eleven participants (30.56%) mentioned that listening instruction was acceptable. No participant classified listening instruction as poor or very poor. In addition, 17 (47.22%) participants considered listening instruction important and very important. Only two participants (5.56%) mentioned that listening instruction was moderately important. No participant believed that listening instruction was slightly important or not important. These results show that participants recognize the importance of listening instruction in ESL settings; however, a significant number of participants consider that listening instruction needs improvement.

Additionally, the survey requested that participants evaluate the audio quality and quantity during language classes. Concerning quality, the results were varied. Six participants (16.67%) rated the audio quality as very good, and ten (27.78%) labeled the audio quality and quantity a good. The majority of the participants (n = 11; 30.56%) considered the audios acceptable, while nine (25.00%) mentioned that the audios were poor in quality. No participant ranked them as very poor. In regard to the audio quantity, eight participants (22.22%) believed it was very good. The same number of participants (n = 14; 38.89%) ranked the audio quantity as good or acceptable. No participant chose poor or very poor for this section. These results demonstrate that the use of audio should be revised, especially in terms of quality.

Finally, participants were asked about the challenges they faced when listening to the audios. The first question asked participants if the overall audio quality (background noise or music, static, etc.) had ever increased the difficulty level of an audio exercise in language classes at the university. Most participants answered affirmatively (n = 28; 77.78%). Only two participants (5.56%) mentioned that overall audio quality had not been an issue. Six participants (16.67%) did not remember any event where overall audio quality had been an issue. The second question asked participants whether the speaker's voice (accent, volume, speed, etc.) had ever increased the difficulty level of an audio exercise in language classes at the university. Most participants answered affirmatively (n = 29; 80.56%). Five participants (13.89%) answered that this has never been an issue. Only two participants (5.56%) claim to not remember any instance where the voice quality hindered their understanding. The findings demonstrate that participants do not always consider that audios are appropriate for ESL settings.

The strengths and weaknesses of each audio were a recurring theme in the responses to the optional, open-

ended question. The following examples summarize participants' opinions concerning the audio.

> Example 3. It sounds kind of robotic sometimes, but it's acceptable enough. (Participant 6, Female AI voice)
>
> Example 7. The audio is clear but the voice is too robotic and does not sound natural. (Participant 14, Male AI voice)
>
> Example 13. It's very pleasant, however, it's (*sic*) feels rushed and even though it certainly has emotion, it's not necessary (*sic*) to be overlly (*sic*) happy nor too excited. It's very fluent yet it feels like some air is neccesary (*sic*) in order to continue the reading. Pretty good though. (Participant 20, Female human voice)
>
> Example 19. Speed was a bit quick for a short story, maybe a little bit of excitement would be good. (Participant 33, Male human voice)

In general, participants also commented on the quality of the headphones used. According to the participants' comments, they were very pleased with the equipment. The equipment used during listening instruction was beyond the scope of this study; however, it should be considered in future research on listening instruction or AI voiceovers.

This investigation shed light on how students perceive human and AI voices. It also discussed the different criteria used to rank voices in ESL environments. Finally, it described students' perceptions of audio use and instruction.

DISCUSSION

Limitations and Future Directions

This study has three main limitations worth mentioning. First, the number and type of criteria used are limited. Only ten criteria were used, and other options could have been considered for this study. However, due to time constraints, the ten most relevant choices, according to the researchers' pilot test, were included. Other or more criteria could trigger different results. Second, the researcher used four types of audio. The selection was based on the results of the pilot test for AI voices using one specific piece of software. Other software may include voices that are more appealing to students or have a more remarkable resemblance to human voices. On the part of human voices, the researcher used professional voiceover experts. They have the necessary equipment and record in a professional studio. Although this was done with the intention of replicating the null environment of AI voices, not all audios used in ESL classes share similar characteristics. Finally, AI audios can be manipulated. In the present study, the audios were not manipulated to standardize procedures. However, using the source software or a third-party application, modified audio may improve AI audios. These modifications may also alter the results from one study to the next. Therefore, the results included in this study cannot be generalized but should serve as a base for future research.

Researchers should replicate this study in other ESL settings or other types of TTS software. For example, not all institutions or professors may have access to the same software. Although TTS free software exists, its quality and number of available voices may not compare to paid software. In addition, future research should consider other types of environments in which noise and background noise may play a part in regular listening instruction. Further research should also determine whether other characteristics or criteria may trigger other results.

CONCLUSIONS

Although this study's findings are not generalizable beyond the study sample, several conclusions can be drawn from the analysis of the results. First, AI voices are not yet at the same level as human voices. In general, human voices are preferred over human voices; however, this does not imply that AI voices should not be used. Some students did not notice that they were listening to a non-human voice; even human voices, recorded by experts and with professional equipment, were criticized in some aspects. In addition, AI voices cannot be used in all scenarios and contexts. For example, AI voices are limited since they cannot create role-plays, dialogues, or other interactive communicative instances without a lot of human intervention, at least not with the type of TTS software used. As AI

voices are not as appealing as human voices, they can be used to generate instructions for listening exercises, provide audio support for readings (especially for visually impaired students), give people who have lost their voices for medical reasons the ability to communicate orally, or create introductions or summaries of listening exercises. Finally, AI voices may be modified to play a more pedagogical role by providing extra audio input or audio prompts for students to discuss various topics.

Second, AI voices do not fall behind in all criteria. This information may be useful for two populations. On the one hand, people who program TTS applications may seek to adjust, to the best of current technological capabilities, those characteristics that mark AI voices as non-human. On the other hand, language professors and material developers may take advantage of this information and include AI or human voices according to their specific needs. For example, in lieu of having a human record specific audios for beginners, a professor may decide to use AI voices since students' main challenge is speed, a feature that is easily adjusted in a computer-generated environment. On the other hand, audio that requires enthusiasm, emotion, or varied intonation may be more suitable for human voices. In addition, AI voices may be useful where resources and exposure to real-life languages are limited. Although the Internet is an excellent source for audio input, finding suitable audios for students' specific needs (accent, speed, topic, duration, vocabulary or grammar level, etc.) may be time-consuming or virtually impossible without considering that some audios may be subject to copyright laws.

The results of this study call for a revision of the program's listening instructions. Although students recognize the importance of listening instruction, they perceive some weaknesses in the instruction they receive. In particular, a relevant group of students considers that the number of audios, their quality, and the general quality of instruction are areas that need improvement. The results do not indicate that these areas need to be completely restructured; however, they point to a systematic revision of current policies and materials to provide students with better and more substantial exposure to auditory input. Currently, policies and materials should also be examined to guarantee that students are exposed to audios according to their level and needs. When students perceive that audios pose additional challenges created by static, unnecessary background noise or music, volume, or accent, among other factors, they may develop negative feelings towards listening exercises. Nevertheless, this does not mean that students should not be challenged. Students may face real-life situations where some of these added difficulties are present; however, institutions should develop clear guidelines to provide students with appropriate materials for their level, age, or other conditions.

It is important to remember that users of TTS software, including Speechelo, can adjust pitch level, breathing, speed, and emphasis to make voices sound more natural. Although Speechelo was created with video creators in mind, its use may provide opportunities to improve students' language learning capabilities. The author suggests that other language programs replicate this study to examine other possible TTS software uses or test its possible improvement in the coming years.

BIBLIOGRAPHICAL REFERENCES

Abbott, R. (2020). *The Reasonable Robot: Artificial Intelligence and the Law* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781108631761

Adamopoulou, E., & Moussiades, L. (2020). An Overview of Chatbot Technology. In I. Maglogiannis, L. Iliadis, & E. Pimenidis (Eds.), Artificial Intelligence Applications and Innovations (Vol. 584, pp. 373–383). Springer International Publishing. https://doi.org/10.1007/978-3-030-49186-4_31

Al-Jarf, R. (2022). Text-to-speech software for promoting EFL freshman students' decoding skills and pronunciation accuracy. Journal of Computer Science and Technology Studies, 4(2), 19-30.

Akmajian, A., Farmer, A. K., Bickmore, L., Demers, R. A., & Harnish, R. M. (Eds.). (2017). *Linguistics: an introduction to language and communication* (Seventh edition). The MIT Press.

Anis, M. (2023). Leveraging Artificial Intelligence for Inclusive English Language Teaching: Strategies And Implications For Learner Diversity. Journal of Multi-

disciplinary Educational Research. 12(6). http://ijmer.in.doi./2023/12.06.89

Arora, V. (2022). *Artificial intelligence in schools: a guide for teachers, administrators, and technology leaders*. Routledge.

Bione, T., Grimshaw, J., & Cardoso, W. (2017). An evaluation of TTS as a pedagogical tool for pronunciation instruction: the 'foreign' language context. In K. Borthwick, L. Bradley, & S. Thouësny (Eds.), *CALL in a climate of change: adapting to turbulent global conditions – short papers from EUROCALL 2017* (pp. 56–61). Research-publishing.net. https://doi.org/10.14705/rpnet.2017.eurocall2017.689

BlasterOnline. (2023). Speechelo [Computer software]. Romania. Retrieved from: https://app.blasteronline.com/speechelo/

Bouck, E. C. (2017). *Assistive technology*. Sage Publications.

Brace, J., Brockhoff, V., Sparkes, N., & Tuckey, J. (2006). *Speaking and listening map of development: addressing current literacy challenges* (2nd ed). Rigby-Harcourt EducationRigby.

Brown, H. D., & Lee, H. (2015). *Teaching by principles: an interactive approach to language pedagogy* (Fourth edition). Pearson Education.

Burgess, S., & Head, K. (2005). *How to teach for exams*. Longman.

Calais-Germain, B., & Germain, F. (2016). *Anatomy of voice: how to enhance and project your best voice* (First U.S. edition). Healing Arts Press.

Cameron, R. M. (2019). *A.I. - 101: a primer on using artifical intelligence in education*. publisher not identified.

Cardoso, W., Smith, G., & Garcia Fuentes, C. (2015). Evaluating text-to-speech synthesizers. *Critical CALL – Proceedings of the 2015 EUROCALL Conference, Padova, Italy*, 108–113. https://doi.org/10.14705/rpnet.2015.000318

Celce-Murcia, M., Brinton, D., & Goodwin, J. M. (2010). *Teaching pronunciation: a course book and reference guide* (2nd ed). Cambridge University Press.

Charpentier-Jiménez, W. (2019). University students´ perception of exposure to various English accents and their production. *Actualidades Investigativas En Educación*, *19*(2), 1–27. https://doi.org/10.15517/aie.v19i2.36908

Chen, L. W., Watanabe, S., & Rudnicky, A. (2023). A vector quantized approach for text to speech synthesis on real-world spontaneous speech. arXiv preprint arXiv:2302.04215.

Cook, A. M. (2019). *Assistive technologies: principles and practice* (5th edition). Elsevier.

Craig, S. D., & Schroeder, N. L. (2019). Text-to-Speech Software and Learning: Investigating the Relevancy of the Voice Effect. *Journal of Educational Computing Research*, *57*(6), 1534–1548. https://doi.org/10.1177/0735633118802877

Dell, A. G., Newton, D. A., & Petroff, J. G. (2017). *Assistive technology in the classroom: enhancing the school experiences of students with disabilities* (Third edition). Pearson.

Derwing, T. M., & Munro, M. J. (2015). *Pronunciation fundamentals: evidence-based perspectives for L2 teaching and research*. John Benjamins Publishing Company.

Dutoit, T. (1997). *An introduction to text-to-speech synthesis*. Kluwer Academic Publishers.

Emiliani, P. L., & Association for the Advancement of Assistive Technology in Europe (Eds.). (2009). *Assistive technology from adapted equipment to inclusive environments: AAATE 2009*. Washington, DC : IOS Press.

Evans, G., & Blenkhorn, P. (2008). Screen Readers and Screen Magnifiers. In M. A. Hersh, M. A. Johnson, & D. Keating (Eds.), *Assistive technology for visually impaired and blind people*. Springer.

Field, J. (2011). Psycholinguistics. In J. Simpson (Ed.), *The Routledge handbook of applied linguistics* (1st ed). Routledge.

Fitria, T. N. (2023). English Accent Variations of American English (Ame) and British English (Bre): An Implication in English Language Teaching. Sketch Journal: Journal of English Teaching, Literature and Linguistics, *3*(1), 1-16.

Green, J. L. (2018). *Assistive technology in special education: resources to support literacy, communication, and learning differences* (Third edition). Prufrock Press, Inc.

Gulson, K. N., Sellar, S., & Webb, P. T. (2022). *Algorithms of education: how datafication and artificial intelligence shape policy*. University of Minnesota Press.

Hadfield, J., & Hadfield, C. (2008). *Introduction to teaching English* (1. publ). Oxford Univ. Press.

Harmer, J. (2007). *How to teach English.* (New ed., 6. impr). Pearson/Longman.

Harmer, J. (2013). *The practice of English language teaching: with DVD* (4. ed., 8. impression). Pearson Education.

Hartono, W. J., Nurfitri, N., Ridwan, R., Kase, E. B., Lake, F., & Zebua, R. S. Y. (2023). Artificial Intelligence (AI) Solutions In English Language Teaching: Teachers-Students Perceptions And Experiences. Journal on Education, *6*(1), 1452-1461.

Hersh, M. A., Johnson, M. A., Keating, D., & Hoffmann, R. (Eds.). (2008). Speech, Text and Braille Conversion Technology. In *Assistive technology for visually impaired and blind people*. Springer.

Hillaire, G., Iniesto, F., & Rienties, B. (2019). Humanising Text-to-Speech Through Emotional Expression in Online Courses. *Journal of Interactive Media in Education*, *2019*(1), 12. https://doi.org/10.5334/jime.519

Holmes, J. N., & Holmes, W. (2001). *Speech synthesis and recognition* (2nd ed). Taylor & Francis.

Honorof, D., McCullough, J., & Somerville, B. *Comma Gets A Cure | IDEA: International Dialects of English Archive*. https://www.dialectsarchive.com/comma-gets-a-cure

Jeste, D. V., Graham, S. A., Nguyen, T. T., Depp, C. A., Lee, E. E., & Kim, H.-C. (2020). Beyond artificial intelligence: exploring artificial wisdom. *International Psychogeriatrics*, *32*(8), 993–1001. https://doi.org/10.1017/S1041610220000927

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. https://doi.org/10.1007/s10772-009-9039-3

Karpf, A. (2006). *The human voice: how this extraordinary instrument reveals essential clues about who we are* (1st U.S. ed). Bloomsbury Publishing.

Kent, D. (2022). *Artificial intelligence in education: fundamentals for educators*. Kotesol DDC.

Kindersley, D. (2023). *Simply Artificial Intelligence*. DK PUBLISHING.

King, M. R., & chatGPT. (2023). A Conversation on Artificial Intelligence, Chatbots, and Plagiarism in Higher Education. *Cellular and Molecular Bioengineering*, *16*(1), 1–2. https://doi.org/10.1007/s12195-022-00754-8

Kochmar, E. (2022). Getting started with Natural Language Processing. Manning Publications.

Kumar, Y., Koul, A. & Singh, C. (2023). A deep learning approaches in text-to-speech system: a systematic review and recent research perspective. Multimed Tools Appl 82, 15171–15197 https://doi.org/10.1007/s11042-022-13943-4

Luo, B., Lau, R. Y. K., Li, C., & Si, Y. (2022). A critical review of state□of□the□art chatbot designs and applications. *WIREs Data Mining and Knowledge Discovery*, *12*(1). https://doi.org/10.1002/widm.1434

McRoy, S. (2021). *Principles of natural language processing*. Susan McRoy.

Memon, S. A. (2020). *Acoustic Correlates of the Voice Qualifiers: A Survey* (arXiv:2010.15869). arXiv. https://doi.org/10.48550/arXiv.2010.15869

Mitchell, M. (2019). *Artificial intelligence: a guide for thinking humans*. Farrar, Straus and Giroux.

Moybeka, A. M., Syariatin, N., Tatipang, D. P., Musthoza, D. A., Dewi, N. P. J. L., & Tineh, S. (2023). Artificial Intelligence and English Classroom: The Implications of AI Toward EFL Students' Motivation. Edumaspul: Jurnal Pendidikan, 7(2), 2444-2454.

Narayanan, S. S., & Alwan, A. (Eds.). (2005). *Text to speech synthesis: new paradigms and advances*. Prentice Hall Professional Technical Reference.

Nass, C. I., & Brave, S. (2005). *Wired for speech: how voice activates and advances the human-computer relationship*. MIT Press.

Nation, I. S. P., & Newton, J. (2009). *Teaching ESL/EFL listening and speaking*. Routledge.

Norton, B., & Toohey, K. (2011). Identity, language learning, and social change. *Language Teaching*, *44*(4), 412–446. https://doi.org/10.1017/S0261444811000309

Patel, M. F., & Jain, P. M. (2008). *English language teaching: (methods, tools & techniques)*. Sunrise Publishers & Distributors.

Paz, K. E. D. S., Almeida, A. A., Behlau, M., & Lopes, L. W. (2022). Descritores de qualidade vocal soprosa, rugosa e saudável no senso comum. *Audiology - Communication Research*, *27*, e2602. https://doi.org/10.1590/2317-6431-2021-2602

Raaijmakers, S. (2022). *Deep learning for natural language processing*. Manning Publications Co.

Taylor, P. A. (2009). *Text-to-speech synthesis*. Cambridge University Press.

Ur, P. (2012). *A course in English language teaching* (2nd ed). Cambridge University Press.

Wang, C., Chen, S., Wu, Y., Zhang, Z., Zhou, L., Liu, S., & Wei, F. (2023). Neural codec language models are zero-shot text to speech synthesizers. arXiv preprint arXiv:2301.02111.

Watkins, P. (2010). *Learning to teach English: a practical introduction for new teachers* (Reprinted). Delta Publishing.