2021-07-30

Text-to-Speech Systems: Learner Perceptions of its Use as a Tool in the Language Classroom

Joseph Chi Man Mak
Brigham Young University

Text-to-Speech Systems: Learner Perceptions of its Use

as a Tool in the Language Classroom

Joseph Chi Man Mak

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Arts

Troy Cox, Chair
K. James Hartshorn
Grant Eckstein

Department of Linguistics

Brigham Young University

ABSTRACT

Text-to-Speech Systems: Learner Perceptions of its Use
as a Tool in the Language Classroom

Joseph Chi Man Mak
Department of Linguistics, BYU
Master of Arts

Text-to-speech (TTS) systems are ubiquitous. From Siri to Alexa to customer service phone call options, listening in a real-world context requires language learners to interact with TTS. Traditionally, language learners report difficulty when listening due to various reasons including genre, text, task, speaker characteristics, and environmental factors. This naturally leads to the question: how do learners perceive TTS in instructional contexts?

Since TTS allows controls on speaker characteristics (e.g. gender, regional variety, speed, etc.) the variety of materials that could be created—especially in contexts in which native speakers are difficult or expensive to find—makes this an attractive option. However, the effectiveness of TTS, namely, intelligibility, expressiveness, and naturalness, might be questioned for those instances in which the listening is more empathic than informational.

In this study, we examined participants' comprehension of the factual details and speaker emotion as well as collected their opinions towards TTS systems for language learning. This study took place in an intensive English Program (IEP) with an academic focus at a large university in the United States. The participants had ACTFL proficiency levels ranging from Novice High to Advance Low. The participants were divided into two groups and through a counterbalanced design, were given a listening assessment in which half of the listening passages were recorded by voice actors, and other half were generated by the TTS system. After the assessment, the participants were given a survey that inquired their opinion towards TTS systems as learning tools.

We did not find significant relationships between the voice delivery and participants' comprehension of details and speakers' emotions. Furthermore, more than half of the participants held positive views to using TTS systems as learning tools; thus, this study suggested the use of TTS systems when applicable.

Keywords: listening, text-to-speech, material development

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Introduction

Text-to-speech (TTS) systems are ubiquitous. From Siri to Alexa to customer service phone call options, listening in a real-world context requires language learners to interact with TTS. This naturally leads to the question: how do learners perceive TTS in instructional contexts?

Listening plays an essential role in communication, both in daily and academic settings. However, learners have reported various factors that cause listening comprehension difficulty. Some common factors of comprehension difficulty can be related to contexts and types of listening tasks, speakers' characteristics, listeners' attitudes, and environmental factors (Goh, 1999). Given the factors of different difficulties, TESOL practitioners and scholars have spent extensive effort on this subject to enhance students’ overall listening comprehension. To decrease the difficulties of listening, we believe TTS systems could be a good source of language input since they require relatively low cost and allow more control on the production by manipulating the variety, accent, rate of speech, and other factors. Thus, the purpose of the study is to assess the influence and effectiveness of listening passages produced by TTS systems and investigate whether TTS systems can be used as appropriate replacements for traditional listening passages made with human voices.

Literature Review

Common Challenges for Learners

Every language learner encounters difficulties in listening to the target language. Some of these difficulties include attention failure the speaker characteristics (Goh, 2000), the quality, and the types of listening (Hasan, 2000). For instance, if learners are only familiar with a particular variety of the language, different accented speech of the same language might negatively impact their comprehension. Furthermore, Hasan (2000) suggested listening that involves interactions, such as conversational listening, requires more effort as it is relatively more difficult than simple informational listening in which listeners only need to understand the factual details. Goh's (2000) research supported this assertion by finding that participants reported that they could understand words but not the intended message. Learners' interpretations could be influenced by the speakers' tone of voice, changes in rate of delivery, word stress and emotion. Since TTS allows controls on speaker characteristics (e.g. gender, regional variety, speed, etc.) the variety of materials that could be created—especially in contexts in which native speakers are difficult or expensive to find—makes this an attractive option for straightforward messages that communicate factual details. However, how might TTS be perceived when used in situation beyond factual details?

Empathic Listening

TTS systems have been used to increase listening comprehension of informational detail, but only a few studies investigated the effectiveness of TTS systems when produced empathic listening. Different from comprehensive listening and critical listening which emphasize understanding of central ideas and persuasive messages, empathic listening focuses on speakers' emotions within a dyadic context (Arnett & Nakagawa, 1983). In different forms of communication, empathic listening plays a fundamental role to help maintain a healthy interpersonal relationship. Thomlison (1991) stated that the core of empathic listening is feeling with another person. Listening empathically requires understanding the speakers' intended messages as well as being sensitive to their emotional state. He included three basic components of empathic listening: 1) relating to our partner's world of experience, 2) recognizing feelings and content in the message, and 3) responding with active feedback. Some studies refer to empathic details as the expressiveness of the speech, which speech characteristic TTS systems cannot fully generate (Lai & Wood, 2000; Lai et al., 2001).

As many studies have illustrated the different factors that influence listening comprehension, some scholars (Arnett & Nakagawa, 1983; Thomlison, 1991) attempted to shift the focus from comprehension of the factual details to comprehension of the speakers' emotions and purposes, which is called empathic listening. In other words, when assessing listening comprehension, language instructors should check learners' comprehension of both factual and empathic details.

Challenges for Teachers

Creating materials can be difficult for teachers and curriculum creators. Authentic materials serve as effective and beneficial tools in language learning; however, it is difficult to find them from either printed or online resources (Sha, 2010). If practitioners were to create their own listening activities, they would have to spend extensive time and money hiring voice actors and adjusting audio when necessary. To reduce the time and cost, therefore, TTS systems might be a practical option.

Text-to-speech (TTS) systems

TTS systems are computer programs that convert written texts to audio files. While originally developed to assist the blind(Handley, 2008), the technology soon entered the main stream. Several TTS systems were introduced to the field since their development and generally consists of a two-step procedure. First, the text is first converted to an abstract underlying linguistic representation. Then, the sequence of phonemes and other linguistic codes is converted to sound by a set of rules (Klatt, 1982).

MITalk (1980) and Klattalk have been used and evaluated to convert English text to natural speech for decades. Below we evaluate TTS systems and the extent to which they prevent or enhance learners' listening comprehension.

Quality of TTS systems

While using a TTS system can efficiently control and lower the difficulties caused by the environmental factors, such as reducing the background noise and increasing the sound quality, it does not eliminate difficulties created by unfamiliar registers or discourses of the text.

To minimize some of the negative impacts, many TTS systems (Klatt, 1982; Pisoni & Hunnicutt, 1980) use an "exceptions" dictionary to correctly produce exceptional common words that would be incorrect if it followed the letter-to-phoneme rules (1982). In addition, when these systems encounter unstressed function words, such as "and," "of," and "the," etc., the system does not handle them with the letter-to-sound rules. As a result, the system can produce more authentic and accurate speech. When MITalk was evaluated, Pisoni and Hunnicutt (1980) stated that modified rhyme test "showed an average error rate of only 6.9% overall." They concluded that it "can produce not only highly intelligible and natural-sounding speech but the quality of the spoken output can be understood and comprehended by native listeners at reasonably high levels of those with only a small amount of listening experience" (p. 574).

Human reaction experiment with TTS

Ever since TTS systems came out, scholars and researchers (Bione et al., 2016; Cardoso et al., 2015; Lai et al., 2001; Sha, 2010; Viswanathan & Viswanathan, 2005) have been interested in their communicative effectiveness. The first and foremost concern regarding TTS systems has always been intelligibility, naturalness, and expressiveness. Some findings (Hjalmarsson, 2011) indicate different listening performance and comprehension levels of users between TTS voices and human speech. Furthermore, some participants reported that TTS voices sounded unnatural and were unpleasant to listen to (Lai & Wood, 2000). These historic disadvantages have been the major drawback of TTS systems.

Lai and Wood (2000) generated materials with five different TTS engines, and they found no significant difference in comprehension of TTS speech among the five engines used. The participants were required to listen to TTS speech in their first language and answer some comprehensive questions. The study suggested that there are no significant gaps between different TTS systems in terms of their sound quality and expressiveness. However, this assumption only indicates the improvement of the engines but does not show the possibility of replacing human voices with TTS voices. With different variables, research subjects have demonstrated similar patterns of performance in listening tasks delivered by TTS voices. Some participants claimed that using TTS voices for reading e-mail would be fine but not for longer items, such as news articles. In conclusion, the participants, even though the listening passages were generated in their first language, were not in favor of TTS voices even if they were allowed to take notes (Lai & Wood, 2000) or the length and complexity of the verbal messages were altered (Lai et al., 2001).

More negative impacts of TTS voices have been examined by other researchers. Hjalmarsson (2009) suggested the number of turn-taking cues, intonation, phrase-final lengthening, semantic completeness, and lip-smacks might affect listeners' judgements. Furthermore, TTS voices may fail to indicate non-verbal behaviors. Although this study suggested that there was no difference in reaction times between these two conditions (the messages were delivered with TTS voices or human voices), there were still some observable differences in participants' judgement distribution.

The final note in this section is that the sound quality and naturalness have not been studied in the field of language learning. Native speakers have been eager to enhance the effectiveness of TTS voices. Researchers and material developers have attempted to find the causes and their respective solutions to improve TTS systems. For instance, Higginbotham et al. (1994) found that improved performance occurred when synthetic speech was presented at a slower rate compared to those presented at a normal rate.

Although studies mentioned in this section primarily focus on native speakers' initial reaction TTS voices, Viswanathan M. and Viswanathan M. (2005) measured speech qualities of TTS systems and their impact on English language learners. They assessed the naturalness of the speech by examining learners' ease of listening, pleasantness, and audio flow as well as the intelligibility by reviewing learners' ratings of the comprehension, pronunciation, articulation, and speaking rate of the TTS speech. The result indicated the system scored better on intelligibility than on naturalness.

Using TTS systems as a tool in the language classroom

Creating listening materials has been a challenging task for language teachers and material designers for a few reasons; there may be constraints related to available budget and time. Some research (Sha, 2010) reported issues with generating listening materials and assessments due to unavailability of native voice talent and editing equipment. In addition, some teachers reported the difficulty in making resources that fit in a specific class or purpose. Because of the difficulties of creating and editing listening materials, the improvement of TTS systems provides an alternative for language teachers. Sha (2010) summarized the advantages of TTS speech over conventional speech tape recording. In his words, TTS systems 1) allow language teachers far more flexibility and adaptability in authoring audio materials; 2) offer highly precise and adjustable speech rate control; 3) generate test items that can be assigned to different voices and in a single audio file; 4) produce files that are easy to copy and distribute; and 5) are more cost-effective than pre-recording human actors.

Given the greater flexibility and variety in language teaching and learning, TTS systems have been used in many classrooms. Most teachers and researchers found TTS systems beneficial to learners' acquisition of writing, vocabulary and reading, and pronunciation (Bione et al., 2016; Cardoso et al., 2015) since TTS systems increase language input.

Of course, the concerns towards TTS systems mentioned in the previous section have carried on in the use of language teaching as well. Cardoso et al. (2015) asked participants to rate the TTS voices at two levels: speech quality and linguistic form. During task 1, they were asked to rate the comprehensibility, naturalness, accuracy, and intelligibility of the voices. During task 2, they were asked to identify the presence of a target feature (English regular past -ed). They found that TTS voices were rated significantly lower than the human-produced samples for all four categories of speech quality. Surprisingly, however, whether the participants were given TTS samples or human-produced samples, they demonstrated almost identical performance in task 2. Such a result further intensifies the argument about whether TTS systems are ready to be used in the language classrooms.

Furthermore, not all research suggests the negative impact on learners' comprehension. Pellegrini et al. (2012) claimed that previous research did not show the whole picture of TTS systems. Instead of testing subjects' comprehension of isolated words, he implemented the co-articulation effect in his dictation experiment. Participants were asked to transcribe a listening passage. All errors in the responses were later identified and analyzed, including spelling, word forms, and contractions. By comparing the use of pre-recorded speech and TTS speech in the dictation task, Pellegrini et al. claimed that the TTS voices were much easier to transcribe than the human speech.

Despite few language learners being in favor of TTS voices in terms of comprehensibility when compared to the human voice, researchers still invested extensive effort to answer whether TTS voices are ready to be used due to the low-cost and flexibility. For instance, Handley (2009) investigated the quality and readiness of different TTS systems in French. He reported that the majority of French TTS systems did not meet the requirements of naturalness and expressiveness. However, he further emphasized that language learners will soon enjoy the benefits from the support of an "untiring non-judgmental substitute native speaker," namely TTS engines.

In other words, a few key indicators must be investigated to determine whether TTS systems are ready for language learning. Although research has suggested that TTS voices do not negatively impact learners' listening comprehension, it was essential to examine learners' comprehension about both factual and empathic details. Furthermore, learners' opinion towards the systems should be considered since listeners' attitudes would influence listeners' comprehension (Goh, 2000). Thus, besides investigating learners' comprehension of the factual details (Bione et al., 2016; Cardoso et al., 2015) and their rating of the intelligibility of TTS voices (Viswanathan M.& Viswanathan M., 2005), this study also focused on how participants viewed the speech expressiveness, namely empathic comprehension. The research questions of this study were summarized as shown below.

Research Questions:

1. How do synthetic voices generated with a text-to-speech (TTS) system influence L2 learners' listening comprehension?

2. To what extent do learners perform differently with TTS listening passages that are informational versus empathic?
3. What is the learners' perception of TTS as a listening tool for language instruction?

Method

Participants

There were seventy-seven participants (N=77; Male=36: Female=41) whose ages ranged from 18-24. Learners with different L1 backgrounds participated in this study with the top four languages being Spanish, Japanese, Portuguese, and Mandarin Chinese (See Table 1). This research took place in an intensive English Program (IEP) with an academic focus at a large university in the United States. The seven levels in this IEP ranged with a rough ACTFL proficiency level equivalency from Novice Low to Advance Low, the students being fairly evenly distributed across the top five levels (see Table 1).

The computer program randomly divided the participants into two groups (A &B) as shown in Table 1.

Table 1

Frequency of L1 and Level

		Level					
Group	**First Language (L1)**	**3**	**4**	**5**	**6**	**7**	**Total**
1	Chinese	0	1	0	0	0	1
	Japanese	2	2	1	1	0	6
	Portuguese	0	1	0	1	2	4
	Spanish	5	4	7	5	4	25
	Other	0	1	1	1	1	4
	Total	7	9	9	8	7	40
2	Chinese	0	1	0	2	1	4
	Japanese	1	1	1	1	0	4
	Portuguese	0	1	0	0	1	2
	Spanish	5	6	2	4	3	20
	Other	1	0	4	1	1	7
	Total	7	9	7	8	6	37
Total	Chinese	0	2	0	2	1	5
	Japanese	3	3	2	2	0	10
	Portuguese	0	2	0	1	3	6
	Spanish	10	10	9	9	7	45
	Other	1	1	5	2	2	11
	Total	14	18	16	16	13	77

Materials

There were two instruments used to gather data in this study: a listening assessment and a brief survey regarding learners' opinions on using TTS systems as a tool in the language classroom. The listening assessment was designed for two purposes of investigating 1) whether TTS voices would influence learners' listening comprehension of factual details, and 2) how learners would rate the speakers' emotions when encountering human voices and TTS voices.

Both instruments were delivered via Qualtrics with the survey being immediately administered after the participants finished the test.

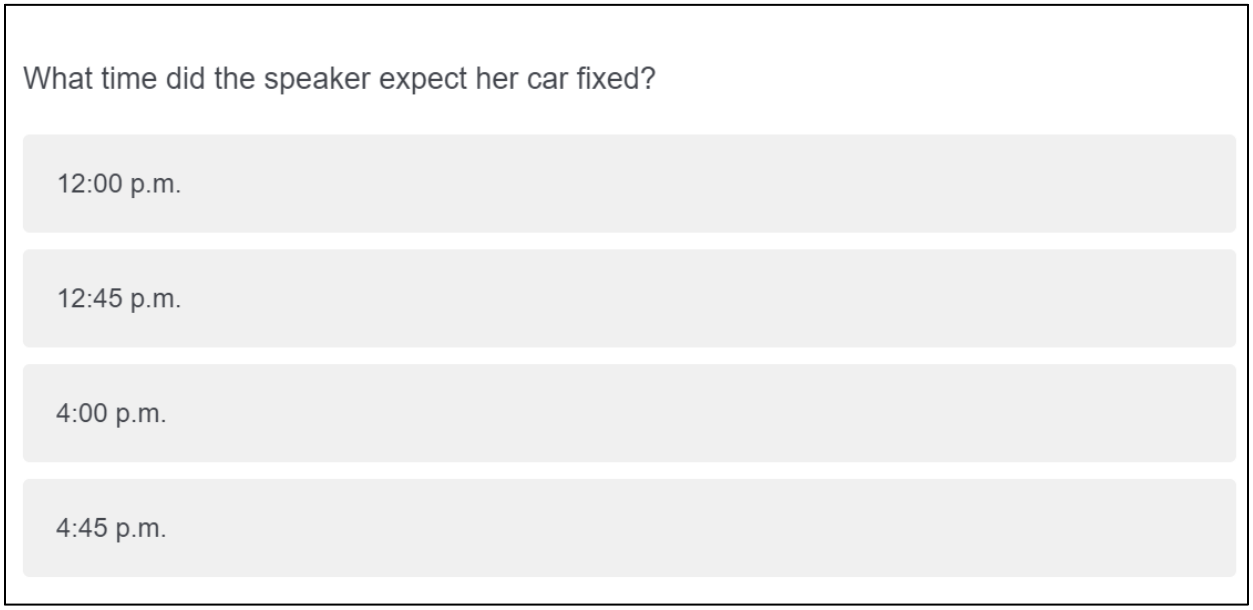
Listening Assessment

The listening assessment contained twelve recordings in the form of voice mails. All listening passages were (1) recorded by human actors and (2) generated by Amazon Polly, a TTS system which grant users flexibility and produces high quality synthesized speech. Each recording lasted no longer than a minute. To ensure the validity of the test, a pilot test was given to another group of learners a semester prior to the official data collection. The original assessment contained eighteen recordings which were recorded and transcribed by three native English speakers; and, after analyzing the validity of the items, the best twelve recordings were chosen, revised, and used in the official data collection.

Three test items followed each recording. The first item was a four-option multiple-choice question about some factual details of the listening (see *Figure* 1). The second item was a short answer question where participants wrote a word or phrase to describe the speaker's emotions, and for the third item, participants rated the speaker's emotion on a 7-point scale (1 = negative, 4 = neutral, 7 = positive) (See *Figure* 2). During the data collection, about half of the participants misunderstood the second the third items regarding speakers' emotion; they gave up on the questions whenever they heard TTS voices in the listening, which behavior was not observed and expected in the pilot test.

Figure 1

Sample MC Test Item



What time did the speaker expect her car fixed?

12:00 p.m.

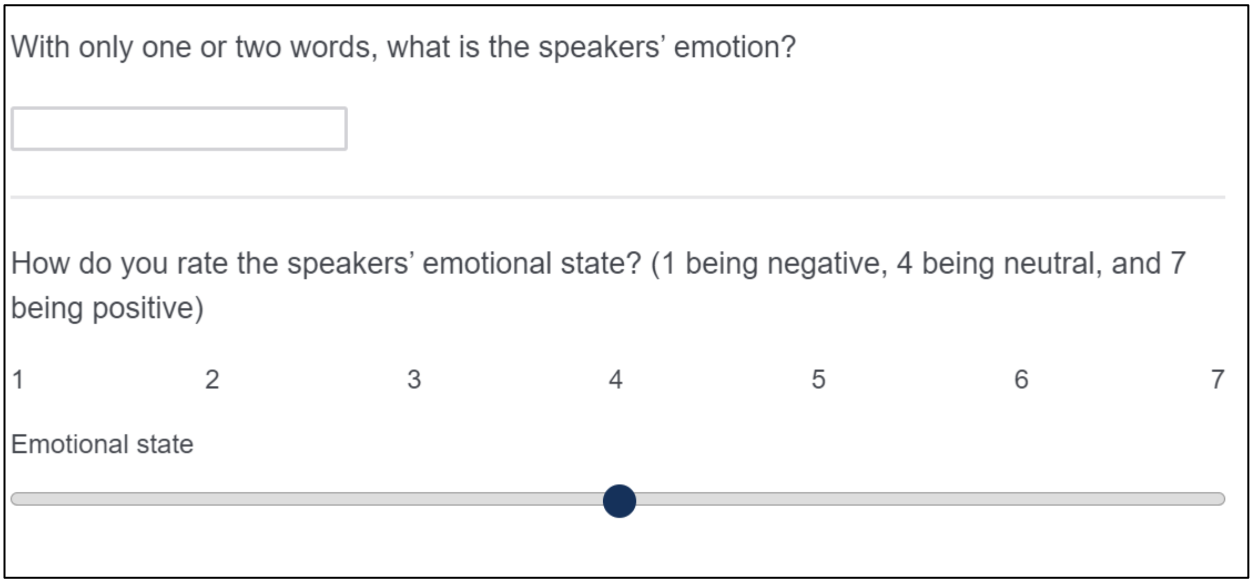
12:45 p.m.

4:00 p.m.

4:45 p.m.

Figure 2

Sample Short Answer and Emotional State Question



With only one or two words, what is the speakers' emotion?

How do you rate the speakers' emotional state? (1 being negative, 4 being neutral, and 7 being positive)

1 2 3 4 5 6 7

Emotional state

Follow-up Survey

In order to answer the third research question of whether students believed a particular delivery method enhanced their listening comprehension. Participants were asked about their experience during the study and their perception towards TTS systems as a tool in the language classroom in a follow-up survey (see *Figure 3)*. All questions in the survey were written in a statement format. Given a scale from one to eight, one being low and eight being high, participants will express the degree of agreement about the statement. Thus, they were able to express their opinions on whether TTS voices influence the comprehension of factual and emotional details. To collect quantitative and qualitative data, at the end of the survey, they were asked to include a short paragraph description regarding their acceptance of using TTS as a tool in learning and teaching.

Figure 3

Questions on the Follow-up Survey

Statements in the follow-up survey

8-point scale questions

1. I can easily recognize when the speaker is human.
2. In terms of accuracy of pronunciation, TTS voices deliver messages as correctly as human voices.
3. In terms of speaking fluency (pause/rhythm/intonation), TTS voices deliver messages as naturally as human voices.

7-point scale questions

4. When the audio is created by a TTS voice, I have increased/decreased comprehension of:
    a. Content
    b. Information
    c. Overall feelings
    d. Speaker's attitudes/emotions
    e. Speaker's intent

8-point scale questions

5. I believe synthetic voices have the potential to enrich students learning experience.
6. I believe synthetic voices have the potential to create good test and homework items.

Open-ended question

7. Would you consider using Text-to-speech systems (TTS) as leaning tools? Explain your answer below.

Research Design

Often in testing, the second time anyone takes a test, their score improved regardless of the instrument. This is called a test ordering effect. To avoid that, we used a counterbalanced design (see *Figure* 4) in which one group answered 6 passages with the human voice first followed by 6 passages with the TTS (Amazon Polly) and the other group did the opposite. If group membership is significant or there is an interaction with group membership, then there is evidence that the test order makes a difference. If there is no interaction, then we can be more confident that the characteristic being evaluated, in this case text entry method as a main effect would account for any differences that were found.

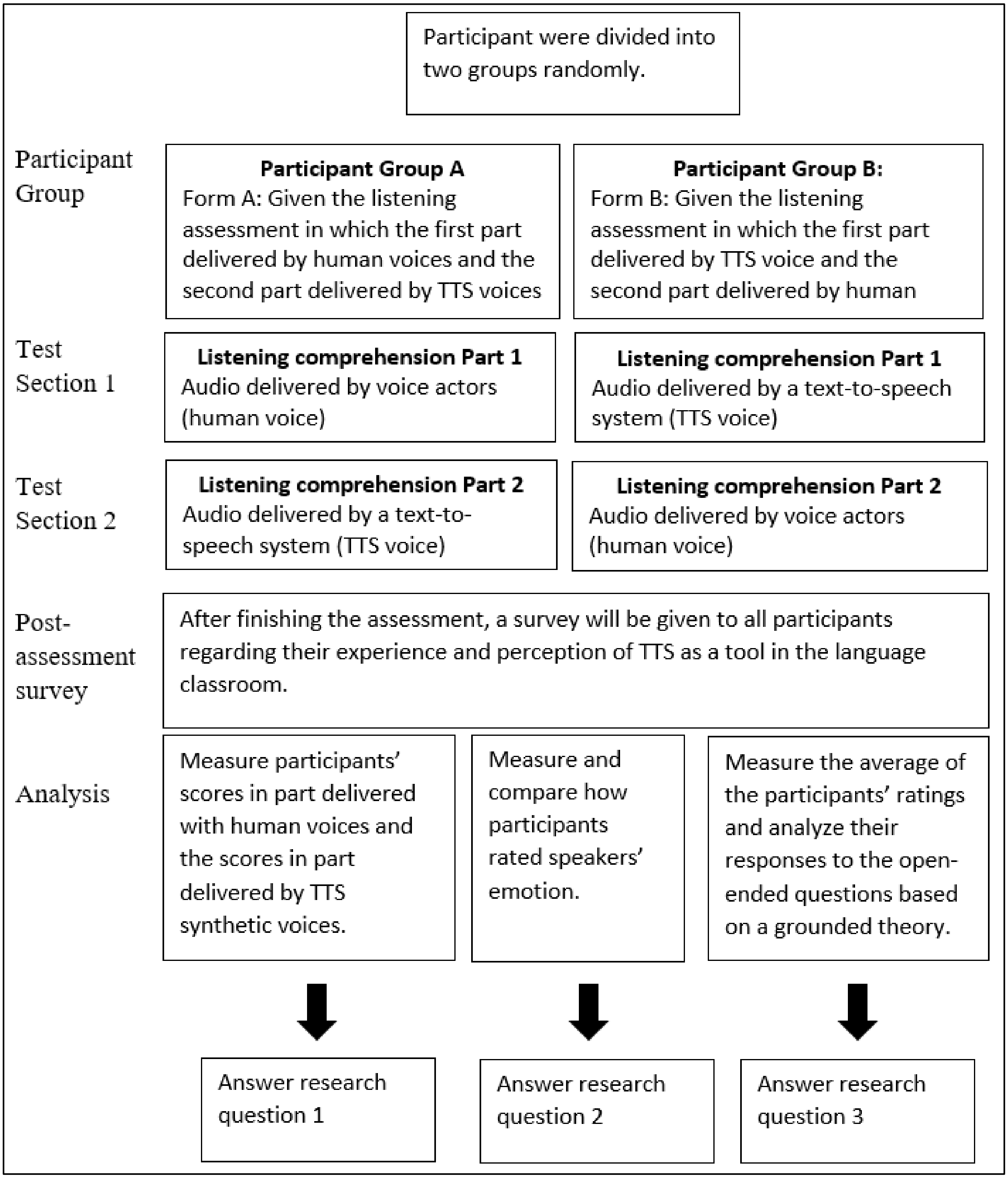
Procedure

The participants completed the instruments during the class time with the teacher as the proctor. Once the participants opened the survey link, they were randomly given either Form A or Form B of the listening assessment. The participants needed to click on the next page after each recording in order to view and answer the test items, and they could not return to listening once they viewed the test items. Participants were asked not take notes during the assessment. Following the assessment, all participants were given a survey regarding the experience and perception of the performance. The assessment took about 25-45 minutes.

In order to ensure the internal consistency and reliability of the assessment, we tried to keep the settings consistent among all participants. All data was collected during the same week in class in either the computer lab or in the classroom with Chromebooks. All teachers received the same proctor instructions as mentioned above. They were guided to prompt students throughout the assessment to avoid problems caused by technological issues, but they were asked not to assist the participants regarding the answers in the comprehensive questions. The instrument had an internal consistency of 0.68 (Guttman's L2 = 0.68) indicating that just under half 46%) of the variability in the listening test scores in attributable to the construct.

Figure 4

Research Procedures



Participant were divided into two groups randomly.

Participant Group

Participant Group A
Form A: Given the listening assessment in which the first part delivered by human voices and the second part delivered by TTS voices

Participant Group B:
Form B: Given the listening assessment in which the first part delivered by TTS voice and the second part delivered by human

Test Section 1

Listening comprehension Part 1
Audio delivered by voice actors (human voice)

Listening comprehension Part 1
Audio delivered by a text-to-speech system (TTS voice)

Test Section 2

Listening comprehension Part 2
Audio delivered by a text-to-speech system (TTS voice)

Listening comprehension Part 2
Audio delivered by voice actors (human voice)

Post-assessment survey

After finishing the assessment, a survey will be given to all participants regarding their experience and perception of TTS as a tool in the language classroom.

Analysis

Measure participants' scores in part delivered with human voices and the scores in part delivered by TTS synthetic voices.

Measure and compare how participants rated speakers' emotion.

Measure the average of the participants' ratings and analyze their responses to the open-ended questions based on a grounded theory.

Answer research question 1

Answer research question 2

Answer research question 3

Data Analysis

To answer the first two research questions that used a counterbalanced design, repeated measures ANOVAs was used. For the first question, the dependent variable was the score on each test section; the within subject independent variable was voice deliver (TTS or human) and the between-subject variable was the group to which participants were randomly assigned. For the second question, the dependent variable was the score on each test section by passage type (informational or empathic); the within-subject independent variable was voice delivery (TTS or human) and the between-subject variable was the group to which participants were randomly assigned.

To answer the third research question, the results of the follow up survey were analyzed. For the first six questions that used scales, we looked at the average rating of the participants' responses to learn more about their perceptions toward TTS systems. For the open-ended question, grounded theory was used to avoid placing preconceived notions on qualitative data It allowed us to examine data from multiple vantage points to help us arrive at a complete picture of the phenomena under investigation. The open-ended question at the end of the survey examined participants' beliefs and attitudes toward TTS systems as tools in the language classroom. Four current ESL teachers examined the responses and consented to use the four categories in dividing the learners' response. When there was disagreement, the most selected option was chosen.

Results

Throughout this study, we investigated three things: how text-to-speech (TTS) systems influence L2 learner' listening comprehension, how learners perform differently with different types of listening passages when the audio is delivered by TTS voices; and the learners' opinions of TTS systems as listening tools in the language classroom. We used different instruments to measure their reactions and performance under different input of voice delivery and conditions (informational and empathic).

RQ 1: Influence of TTS

Voice delivery appeared to have no significant impact on listening comprehension. To examine the effect of voice delivery to English learners, we compared the scores of parts 1 and 2 of the test based on voice delivery. In part 1, group A' scores (*M* = 0.76) were like group B's (*M* = 0.77); in part 2, group A (*M* = 0.84) and group B (*M* = 0.85) had a similar performance as shown in Table 2.

Table 2

Scores in Listening Comprehension Tasks

				95% Confidence Interval	
Group	**Testing section**	**Mean**	**SD**	**Lower**	**Upper**
A (Human>TTS) 40	Part 1 (human)	0.76	0.176	0.707	0.817
	Part 2 (TTS)	0.84	0.155	0.786	0.896
B (TTS>Human) 37	Part 1 (TTS)	0.77	0.201	0.715	0.826
	Part 2 (human	0.85	0.158	0.791	0.902

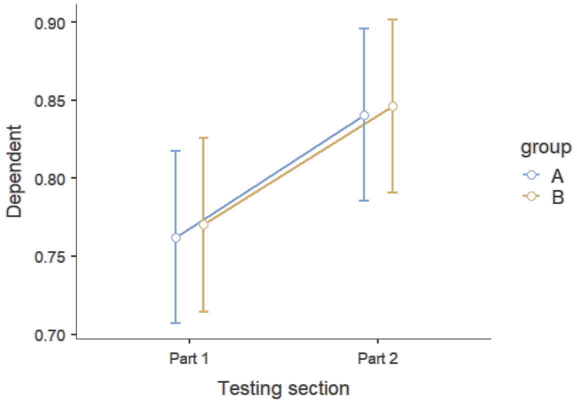


Figure 5. Estimated Marginal Means (Test Sections)

A Repeated measures ANOVA found no significant relationship between any of the variables related to group as a main effect [*F* (1,75) = 0.039, *p* = 0.84]. Thus, the order in which examinees responded to the voice delivery system did not result in any systematic difference. There was a significant difference with test section as a main effect [*F* (1,75) = 17.08, *p* < 0.001] indicating that Part 1 was more difficult than Part 2 regardless of the voice delivery. However, the interaction between test section and group was not significant [*F* (1,75) = 0.003, *p* = 0.95]. Participants had almost identical test scores in each section whether the listening passages were delivered by human actors or TTS systems (see *Figure* 5).

RQ 2: Passage Type (informational vs. empathic)

Passage types together with voice delivery appeared to have no significant impact on listening comprehension either. We measured learners ' scores within four categories for further investigation regarding the effects by passage types (informational or empathic) and voice delivery methods (humans & TTS) on learners’ comprehension (see Table 3). Participants could receive up to 3 points in each category; in general, they performed better with empathic listening questions, whether they encountered human empathic (*M* = 2.64) or TTS empathic (*M* = 2.61).

When presented with informational listening questions, their performance was lower but similar, human informational (*M* = 2.18) and TTS informational (*M* = 2.23).

We used another repeated measure ANOVA and did not find a significant relationship in the within subject variable of groups as a main effect [*F* (1,75) = 0.034, *p* = 0.85] and thus the groups were considered equivalent. We also found a significant difference in the passage types as a main effect [*F* (1,75) = 31.94, *p* < 0.001, partial eta square = 0.30] with a significant Tukey post hoc test [*t* = 5.65, *p* < .001]. In other words, empathic questions were easier than the informational questions. Furthermore, there was not significant difference found in the voice delivery as a main effect [*F* (1,75) = 0.005, *p* = 0.95].

Table 3

Scores in Different Section

					95% Confidence Interval	
Group	**Voice Delivery**	**Passage Types**	**Mean**	**SE**	**Lower**	**Upper**
A (Human/TTS)	Human	Informational	1.97	0.112	1.75	2.20
		Empathic	2.60	0.112	2.38	2.82
	TTS	Informational	2.42	0.112	2.20	2.65
		Empathic	2.62	0.112	2.40	2.85
B (TTS/Human)	Human	Informational	2.41	0.115	2.18	2.63
		Empathic	2.68	0.115	2.45	2.90
	TTS	Informational	2.03	0.115	1.80	2.25
		Empathic	2.59	0.115	2.37	2.82

RQ 3:

Participants seemed to have a positive outlook on the inclusion of TTS in instructional material. Once the participants finished the assessments, the computer program directed them to a follow-up survey regarding their thoughts of using TTS systems as learning tools. The follow-up survey was designed to collect quantitative as well as qualitative data.

Quantitative Data

There were two parts in the follow-up survey; the questions in the first part focused on participants' opinion on the ease of recognizing TTS voices and other factors regarding the quality of delivery on a scale from one (disagree) to eight (agree).

In comparing TTS voices and human voices, learners reported a positive view (equal or greater than 5) on the voice recognition, accuracy, and fluency as shown in Table 4. Even though participants could relatively easily recognize when the listening passages were delivered by TTS voices (*M* = 6.58), they believed TTS voices in the assessment were as correct (*M* = 5.68) and as natural (*M* = 5.18) as human voices. In addition, when the participants were asked if TTS systems have the potential to enrich students' learning experience (*M* = 5. 16) and if they are beneficial for creation of homework and test items (*M* = 5.38), they also expressed a slightly positive view (see Table 4).

Table 4

Opinion on Recognizing TTS Voices (8-point scale)

	Voice recognition	**Accuracy**	**Fluency**	**Enriching EXP.**	**Items Creation**
M	6.58	5.68	5.18	5.16	5.38
SD	1.66	1.85	2.11	1.87	1.99

In the second part of the survey, they were asked to rate their comprehension of the content, information, speakers' feelings, attitude, and intent on a scale from one (disagree) to seven (agree). Four on the scale was a neutral option for participants; it means the participants believed TTS voices neither increase nor decrease their understanding of the listening (See Table 5). When rating whether TTS voices increased or decreased the participants' understanding of the content (*M* = 5.0) and other informational details (*M* = 5. 18), they reported a slightly positive view (means greater than 4). Nevertheless, participants reported a slightly negative view (mean less then 4) when they rated the delivery of feelings (*M* = 3.69), speakers' attitude (*M* = 3.68), and intent (*M* = 4. 23).

Table 5

Opinion on Comprehension (7-point scale)

	Content	Information	Feeling	Attitudes	Intent
Mean	5.00	5.18	3.69	3.58	4.23
SD	1.22	1.35	1.72	1.95	1.63

Qualitative Data

Following the rating, the participants were asked to give a short response to the following questions: Would you consider using Text-to-speech systems (TTS) as learning tools? As mentioned above that four current ESL teacher put all responses were put into four categories which have been thoroughly discussed and revised based on their feedback. The categories included: (1) yes because of their usefulness, (2) yes but only for certain settings, (3) no but may be fine in the future, (4) no because of the difficulty. Some participants did not leave any comments; thus, we could only collect feedback from seventy-two participants in total. Thirty of them held a positive view on TTS systems because of their usefulness; twenty of them believed that TTS systems were useful but only in certain settings. Sixteen participants did not consider using TTS, and only one student expressed his refusal to TTS systems but his potential acceptance in the future.

1. Yes because of their usefulness (n = 30; 42%). The participants found TTS systems useful in assisting learning activities.

- Yes I would use TTS because it is clear to understand. Sometimes human speaks *[sic]* not clear and fast, so sometimes have a hard time understanding human natural speech.
- Yes, I can consider TTS as learning tool because the system *[sic]* is easy to use and very good.
- I like to *[sic]* using Text-to speech systems because if human speaks, they have their features, so it is hard to recognize. But TTS doesn't change its pronounce *[sic]*, so it's better for me.
- I think that TTS is a good learning tool. For me, is a good exam/quiz to pratices *[sic]* my listening skills.

2. Yes but only in certain setting (n = 20; 28%). The participants found TTS systems useful in assisting learning activities, but they also expressed concerns of their application and authenticity. Some participants believed TTS systems are beneficial only in certain conditions, such as studying in an EFL environment.

- It would be helpful for beginner *[sic]* students, sometimes native speakers pronunce *[sic]* different some words, thus, the student might get confused. I believe is really good for begginers *[sic]*, so they can adapt to it within time.
- I think it is good if you don't have any one with you, but if I could chose *[sic]* between a person and TTS. I would prefer a person because it *[sic]* easier for me to understand and pay attention to the speaker. I think the way of people lower of higher there is very important to me.
- Yes, I think sometimes it can be useful because schools, however, about feeling and emotions it will be difficult to understand.

- I consider that the TTS as learning toos *[sic]* is good but it mist *[sic]* reflect the situation of the conversation. The TTS have a good information that could be said correctly and without errors.
- I would consider using TTS as learning tools only if the instituion *[sic]* don't have people for to record. Because it ins't *[sic]* bed TTS, but is better to listen a recording by human voice.

3. No because of their difficulty (n = 16; 22%). The participants denied the statement or found difficulty in understanding the audio due to the change of voice delivery.

- I prefer human's actors because it is more natural, specially if I need to recognize intonations or emotions. I don't [sic] thing that is a good idea to put TTS as a learning tool.
- I would not consider it as learning tools because I can't identify the speker's *[sic]*emotion and I think this is important in a listening test.
- I would not to use TTS. It was so hard to read people feelings. I think, TTS is not useful to people as learning tools, because, language was born by people. Also, language is tool to speak other people, so if we want to improve our language level, we should use the human's voice, because human's voice is the closest to real conversation *[sic]*.

4. No but may use them in the future (n = 1; 1%).

- I think not yet. It needs to improve the intonation and emotion, but could be in the future, but now no.

Conclusion

Summary of Finding

TTS systems did not seem to effect L2 learners’ listening comprehension. This study investigated whether listening passages generated by TTS systems have a positive or negative effect on listening comprehension. Thus, test items that examined participants’ factual comprehension were implemented, and the result did not show a gap in their performance in different sections. Participants had almost identical test scores in different sections no matter they listened to natural or TTS voices. Such a result further confirmed what previous studies have found (Bione et al., 2016; Cardoso et al., 2015).

TTS systems did not negatively influence participants’ comprehension even if given empathic questions. Although Viswanathan M. and Viswanathan M. (2005) suggested learners were more tolerate to the intelligibility than the expressiveness of TTS voices, we found that participants did not have much distinct performance in empathic listening question whether they listened to TTS voices or not. In addition to the subjects' comprehension of the factual details, we examined whether different voice delivery would influence their comprehension of the speakers' intent or emotional state. In this study, half of the questions (*n* = 6) were informational questions, and another half were empathic passages in which speakers' emotional states helped listeners understand the message. As shown in the result, when listening to the empathic passages, participants had similar performances in both sections (generated by natural voices and TTS voices).

Most participants held positive opinions towards TTS systems as a language instrument. Different from what Lai & Wood (2000) found, the participants in this study did not reject TTS systems. Although some of them pointed the drawback of using TTS systems, lack of emotions, they were in favor of TTS systems in assisting learning activities. We collected quantitative and qualitative data concerning participants' views toward TTS systems. Even though participants could recognize when the voices were not human, they also indicated the TTS systems produced speech as accurately and fluently as natural speech; they were also slightly positive in using TTS systems for language learning. However, they stated in both quantitative and qualitative data that they had difficulty interpreting the feelings and attitudes generated by the TTS voices despite the high voice quality. When asked whether they would consider using TTS as a language learning tool, more than half of the participants responded "yes." Only about 25% of the participants said they would not use TTS systems as learning instruments due to the lack of emotion and expressiveness; more importantly, they emphasize that human voices are the main source of listening in all daily settings which TTS systems could not offer.

Limitations

This study had three major limitations that may affected the result of this study: setting, participants' L1, and familiarity with TTS systems. First, this study took place in an intensive English program in the U.S. Besides the language input during class time, learners could access other forms of language input outside the class. Conducting research in such an environment may have influenced how the learners perceived TTS systems as learning tools. Also, more than half of the participants in this study were Spanish speakers, in which their L1 may have a positive or negative effect on listening comprehension. Furthermore, most of their age ranged from 18-24 in which we assumed they were more familiar or accepting of new technologies and online learning tools.

To eliminate these limitations, we suggest increased subject group size and diversity. Also, since learners' attitudes and belief towards TTS systems would influence their comprehension, we suggest shifting the focus from learners' performance when listening to TTS passages to their acceptance of TTS systems as language instruments. As the subject group becomes larger and more diverse, we can determine in what settings TTS systems are ready as learning tools. Moreover, this study only focused on learners' initial reactions to TTS systems. Increased exposure to TTS voices may affect how they perceive the systems as learning tools.

Implications of Findings and Directions for Further Research

Are TTS systems ready for language learning and teaching? If so, language instructors may consider using TTS systems as a cost-effective alternative to create materials and test items. If not, finding, recruiting and recording human voices with the additional time needed would still be the gold standard. Studies have shown that listening is difficult for a few reasons: the texts and genres, speakers' characteristics including environmental factors such as rate and accent and the tasks the listeners need to perform. Although TTS systems cannot alter the difficulty caused by the texts and tasks, they can control speaker characteristics and other environmental factors, by allowing the deliberate manipulation of speed, accent, and other factors to increase or decrease the passage difficulty.

The effectiveness of TTS systems has been questioned by scholars, but we find that the potential makes it worth investigating. Ever since TTS systems were invented and commonly used by people with disabilities and learning difficulties, researchers have been arguing whether language instructors and program administrators should implement them into the learning process. Some studies have suggested that listeners hesitate or need extra time to react when the listeners encountered TTS voices, while other studies have claimed there was an almost identical performance in students whether they listened to natural human voices or TTS voices.

In this study, learners' performance and opinions were examined in response to both delivery methods. Instead of emphasizing learners' response time and effort, this study mainly focused on their comprehension of the content and the speaker's emotion or intent. The result of the assessment was similar to what Cardoso et al. (2015) and Lai and Wood (2000) suggested that there was not much of a difference in participants' performance, and whether they were asked to listen to normal or TTS voice recording, their comprehension of the factual details and the main ideas would not be influenced even if given TTS voices. Since there are libraries of different TTS voices available, then developers can more easily create materials that reflect different varieties including those from other regions of the world as well as variation within a country.

Does it matter whether the texts are informational and empathic? Not as much as one might expect, with informational passages, participants were asked to comprehend the content and details of the listening; with the other type of passages, they were asked to understand the speakers’ intent or emotional state. While understanding the speaker’s emotion plays an essential role in interpreting the message. Interestingly, in the results section, we found the participants performed the same with empathic passages whether they encountered natural or TTS voices. This finding may further suggest the use of TTS systems could be beneficial.

Most participants in this study had optimistic views towards the voice quality of TTS systems for language instructions. To determine if TTS systems were ready for language learning, we collected participants' opinions. Since learners' attitude towards the listening passage or the speaker's voice and accent may affect their performance (Goh, 1999), understanding the learners’ view is essential. Most participants in this study indicated a slightly optimistic view towards the use of TTS systems in language learning, still, some of them expressed concerns about whether lack of emotion and expressiveness would distort their interpretation of the intended message. With the open-ended question, more than half of the participants agreed with using TTS systems on a regular basis especially when studying in an EFL environment in which the access to native speakers is limited. The most common mentioned disadvantage was lack of emotions and expressiveness. While the sound quality and naturalness have been improved, some participants still rejected TTS systems since they didn't believe the tools would be helpful in daily learning activities.

By considering the test result, TTS systems are ready as language learning instruments in many instances. However, based on the opinions and perceptions toward TTS systems collected in the study, TTS systems may not be widely accepted when other resources exist (ESL environment).

Reference

Arnett, R., & Nakagawa, G. (1983). The assumptive roots of empathic listening: A critique. *Communication Education, 32*(4), 368-378. https://doi.org/10.1080/03634528309378558

Bione, T., Grimshaw, J., & Cardoso, W. (2016). An evaluation of text-to-speech synthesizers in the foreign language classroom: Learners' perceptions [Conference paper]. EUROCALL Conference, Padova, Italy. 50-54.

Cardoso, W., Smith, G., & Fuentes, G. C. (2015). *Evaluating text-to-speech synthesizers* [Conference paper]. EUROCALL Conference, Padova, Italy. 108-113.

Goh, C. (2000). A cognitive perspective on language learners' listening comprehension problems. *System, 28*(1), 55-75. https://doi.org/10.1016/S0346-251X(99)00060-3

Handley, Z. (2009). Is text-to-speech synthesis ready for use in computer-assisted language learning? *Speech Communication, 51*(10), 906-919. https://doi.org/10.1016/j.specom.2008.12.004

Higginbotham, D, J., Drazek, A., Kowarshy, K., Scally, C., Segal, E. (1994). Discourse comprehension of synthetic speech delivered at normal and slow presentation rates. *Augmentative and Alternative Communication, 10*(3), 191-202. https://doi.org/10.1080/07434619412331276900

Hjalmarsson, A. (2011). The additive effect of turn-taking cues in human and synthetic voice. *Speech Communication, 53*(1), 23-35. https://doi.org/10.1016/j.specom.2010.08.003

Klatt, D. (1982, May. 3-5). *The klattalk text-to-speech vonversion system* [Conference session]. ICASSP '82. IEEE International Conference on Acoustics, Speech, and Signal Processing, Paris, France. 10.1109/ICASSP.1982.1171431

Lai, J., Cheng, K., Green, P. et al. (2001, Mar, 1). *On the read and on the web? Comprehension* *of synthetic and human speech while driving* [Conference session]. CHI01: Proceedings of the SIGCHI on Human Factors in Computing Systems, Seattle Washington, USA. https://doi.org/10.1145/365024.365100

Lai, J., & Wood, D. (2000, Apr). *The effect of task conditions on the comprehensibility of synthetic speech* [Conference session]. CHI00: Proceedings of the SIGCHI conference on Human Factors in Computing Systems, The Hague, The Netherlands. https://doi.org/10.1145/332040.332451

Pellegrini, T. Trancoso, I. Costa, A. (2012, Sep. 9-13*). Less errors with TTS? A dictation experiment with foreign language learners* [Conference session]. 13th Annual Conference of the International Speech Communication Association, Portland & USA. https://www.isca-speech.org/archive/interspeech_2012/i12_1291.html

Pisoni, D. Hunnicutt, S. (1980, Apr. 9-11). *Perceptual evaluation Of MITalk: The MIT unrestricted text-to-speech system* [Conference session]. ICASSP ;80. IEEE International Conference on Acoustics, Speech, and Signal Processing, Denver, CO, USA. 10.1109/ICASSP.1980.1170888

Thomlison, T. (1991, Mar). *Approaches for teaching empathic listening* [Conference session]. The Annual Meeting of the International Listening Association. Jacksonville, FL, US. https://files.eric.ed.gov/fulltext/ED333489.pdf

Sha, G. (2010). Using TTS voices to develop audio materials for listening comprehension: A digital approach. *British Journal of Educational Technology, 41*(4), 632-641. https://doi.org/10.1111/j.1467-8535.2009.01025.x

Viswanathan, M., & Viswanathan, M. (2005). Measuring speech quality for text-to-speech systems: Development and assessment of a modified mean opinion score (MOS) scale. *Computer Speech and Language, 19*(1), 55-83. https://doi.org/10.1016/j.csl.2003.12.001