Intelligibility of Chinese Synthesized Speech and Learners' Attitudes towards Its Use in CSL Learning and Instruction: A Preliminary Study [1]
(中文合成语音的可理解性及学习者对其在中文二语教学中应用的态度初探)

Wang, Yanlin
(王彦琳)
Texas Tech University
(德克萨斯理工大学)
yanlin.wang@ttu.edu

Da, Jun
(笪骏)
Middle Tennessee State University
(中田纳西州立大学)
jun.da@mtsu.edu

Yin, Chengxu
(尹承旭)
University of Notre Dame
(圣母大学)
chengxu.yin.9@nd.edu

Abstract: Text-to-Speech technology has the potential to produce audio materials for instructional purposes. Before the technology can be used as a technological assistant tool to aid second language instruction and learning, there are second language acquisition and pedagogical questions that need to be addressed. This study investigated the intelligibility of Chinese synthesized speech and CSL (Chinese as a second language) learners' attitudes towards the use of Text-to-Speech technology in second language learning and instruction. Data from a survey with 39 beginning and intermediate level participants from three different American universities showed that CSL learners were able to distinguish between audio recordings made by real persons and those made by computers, and found Chinese synthesized speech intelligible. At the same time, those participants also agreed that Text-to-Speech technology could be helpful and welcomed the use of the technology to assist their Chinese language learning. These findings support the idea that CSL instructors can now experiment with current Text-to-Speech technologies in their classroom instruction so that a better understanding of the effects of Text-to-Speech technology on second language acquisition and issues involved in its instructional use can be achieved.

[1] Results reported in this paper are part of a research project titled *The use of speech processing technologies for beginning and intermediate level CSL (Chinese as a Second Language) learning and instruction* approved by the IRB of Middle Tennessee State University (Protocol ID: 22-1120 2q). Preliminary analysis of results from this study was first presented at the 2022 Chinese Language Teachers Annual Conference by Chengxu Yin, Jun Da, and Yanlin Wang under the title *The use of speech processing technologies for novice- and intermediate-level CSL learning and instruction*.

摘要: 语音合成技术在合成教学听力材料方面具有潜在的应用前景。然而，将其直接用于辅助二语教学之前，有必要深入研究其在二语习得和教学法方面的相关问题。本项研究采用在线问卷探讨两个问题：其一是基于二语学习者视角的中文合成语音的可理解性；其二是在二语习得及教学中使用语音合成技术时，中文二语学习者持何态度。来自三所美国大学的 39 名初、中级中文二语学习者自愿参与了本项研究。问卷数据显示对于真人录制的音频和计算机生成的音频，中文二语学习者能够加以辨别，并能听懂中文合成语音。同时大多数参与调研的学习者均表示语音合成技术有助于提高他们的外语听力技能，支持教师在二语教学中使用语音合成技术。本研究提出现阶段二语教师可以在实际教学中尝试使用语音合成技术，这样可以更好地了解其在二外习得中有哪些效果，以及在教学使用中有哪些问题。

Keywords: Text-to-Speech technology, synthesized speech, intelligibility, learners’ attitudes, learning and instruction of Chinese as a second language

关键词: 文字转语音技术、合成语音、可懂度、学习者的态度、中文二语教学

1. Introduction

Text-to-Speech technology (TTS) refers to the conversion of digital text into speech by computers. Computer-generated speech, or synthesized speech, can be pre-recorded or produced in real time. Since the invention of first general (English) text-to-speech systems in the 1960s (c.f., Wikipedia, n.d.), the technology has gone through several decades of development and matured. It is now readily available as a general-purpose consumer technology and in multiple languages including Chinese, and it can be found on mobile or other smart devices and is embedded in a variety of software applications and services. With easy access to the technology and its capabilities, consumers have adopted it for real world use in a variety of settings. It is not infrequent to find, for example, video clips dubbed with artificial voices on YouTube or other virtual universes such as bilibili where the technology is used in place of real human voices for convenience of production or due to privacy concerns. As TTS technology is becoming ubiquitous and welcomed by consumers, second language learners will most likely encounter it in real-world communications. Accordingly, it is now the time to examine the feasibility of applying the technology for second language learning and instruction.

1.1 Instructional audio materials production in second language learning and instruction

In second language instruction, audio materials used for language development come from two main sources: authentic materials made by native speakers for native speakers' consumption, or customized materials authored by instructors or other speakers for learning purposes. The latter is more often found in beginning and intermediate level classes where learners' language proficiency is limited, and it is more challenging, if not impossible, for instructors to find and adapt authentic materials that would be suitable for this specific group of learners and fit the curriculum design. In the traditional way of producing such instructional audio materials, instructors would record audio materials themselves or find someone else for help. This method of producing audio materials is often time consuming, or sometimes even hard to do because of the lack of personnel (imagine there is only one instructor available to record a conversation) or the lack of availability of the developers.

While the inconvenience and difficulty in recording audio materials by instructors themselves has limited the availability and quantity of suitable listening materials for learners, instructors often receive requests from learners, especially from auditory learners, for additional or supplementary audio materials to reinforce print-based or digital texts for a more robust learning experience. Given the amount of time and labor required to record audio with human voice, these requests are often not met (in full) in practice.

In contrast to recording audio with human voice, computers can generate audio from digital texts easily and efficiently. The convenience afforded by the Text-to-Speech technology makes it an appealing alternative to produce audio materials for second language learning.

1.2 The use of TTS in second language learning and instruction

Text-to-Speech technology was first applied in education as an assistive technology for disabled students such as reading systems for the blind (Taylor, 2009). As an assistive technology, it is mostly used by learners who are native speakers. By enabling auditory input, it helps those learners who would otherwise be deprived of access to learning materials or other inputs. It also helps learners gain or improve other skills such as reading literacy. For example, Wood, et al. (2017) conducted a meta-analysis of prior research on the effectiveness of the technology on reading comprehension and found it modestly effective. When it is used by native speakers, there is less concern if the audio generated by the computer is natural. Robotic speech (sometimes with erroneous prosody and/or pronunciation) is tolerated as long as the synthesized speech is intelligible and does not negatively affect real world communications.

In contrast to the use of Text-to-Speech technology by and for native speakers, when the technology is to be used in second language education, synthesized speech, whether pre-recorded or generated in real time, is most likely to serve as a language role model for second language learners and will interact with learners directly where computers have increasingly become participants in real world communications. Such a functional change brings about a series of second language acquisition and pedagogy questions that have to be addressed if the technology is ever to be applied in second language learning and instruction. For example, is synthesized speech intelligible, and does it sound natural and authentic to non-native speakers? Will second language learners welcome the use of TTS technology for their language learning? In what ways can synthesized speech help second language acquisition?

Prior research has explored some possible applications of Text-to-Speech technology in second language education and its related issues. For example, Bione et al. (2016) reported that EFL (English as a foreign language) learners hold a positive view towards the pedagogical use of TTS, and that they would like to use the technology as a learning tool. Kent (2021) investigated the potential of a voice-user interface in TESOL (Teaching English to speakers of other languages) and reported that participants found it acceptable. Cardoso et al. (2015) reported that synthesized speech was equally effective as human speech when it was used in a listening perception task. Liakin et al. (2017) found that the pedagogical use of mobile TTS technology was helpful in complementing and enhancing the teaching of L2 pronunciation of French liaison.

Text-to-Speech technology is also found to be useful for the development of other second language skills. Huang and Liao (2015) reported that digital materials enhanced with TTS technology helped ESL learners' vocabulary learning and improved their motivation. Proctor et al. (2007) found that the use of TTS reading aloud functionality was associated with vocabulary development and reading comprehension gains.

As compared with prior research on other second languages (mostly ESL, or English as a Second Language), there are, to the best of the authors' knowledge, very limited studies on the use of Text-to-Speech technology for learning Chinese as a second language (CSL). Soon et al. (2020) studied the intrapersonal and interpersonal perceptions of 119 Chinese language learners at a large Malaysian university towards the use of Pinyin TTS system developed by the authors themselves and found that those learners had a modest positive perspective and agreed that the system helped improve their pronunciation. Yeh (2014) conducted a multi-case study of five K-12 Chinese language teachers on the use of Text-to-Speech, Speech-to-Text, and machine translation technologies in their classroom instruction and reported that the attitudes of teachers, administrative support, and ease of use and access to technology were the key factors in the teachers' actual use of the technologies.

Since the sound systems of human languages differ from each other, and speech engines rely on language specific models and parameters to produce audio output, there is the need to (re-)investigate the same issues with the Chinese language and to determine whether similar findings from prior research on other languages such as ESL are also applicable to learning Chinese as a second language (CSL).

1.3 Objectives and research questions of this study

Given the very limited scope and nature of previous research in the case of CSL, this study is designed to investigate two underlying issues that inform whether further exploration of the use of TTS technology for CSL learning and instruction is warranted. The first has to do with intelligibility: Can beginning and intermediate level CSL learners distinguish between audio recordings made by a human and those made by a computer? Can synthesized Chinese speech produced with the current off-the-shelf consumer technology be understood by CSL learners, especially those at the beginning and intermediate levels? Even though the current consensus among native speakers is that synthesized speech is not perfect and its deviation from natural human speech can be easily identified, it is necessary to confirm if second language learners, especially those CSL learners at the beginning or intermediate level, also find synthesized speech intelligible.

The second issue is related to CSL learners' attitudes: What are the attitudes of CSL learners toward the use of synthesized speech for their Chinese language learning? It has long been understood that attitude correlates positively with language learning outcomes (c.f., for example, Gardner, 1968). In the case of using synthesized speech for CSL learning, it will be interesting to explore whether CSL learners hold a positive or negative attitude when they are able to differentiate between human speech and computer-generated speech. A negative attitude will likely affect the adoption and effectiveness of synthesized speech for language acquisition when it is used in learning materials by the instructors or when learners encounter it elsewhere (such as in the digital universe where native speakers are producing increasingly more of such audio materials).

If TTS technology is ever to become a viable alternative for producing audio materials for second language learning or serve as a positive language role model, it is expected that synthesized speech should be equally intelligible to second language learners as to native speakers, and at the same time, be welcomed by second language learners as a technology to aid in their second language learning.

2. This study

This study invited Chinese language learners as volunteers from three American universities to investigate whether CSL learners, especially those at the beginning and intermediate levels, can differentiate between audio recordings made by humans and those made by a computer, and their attitudes towards the use of synthesized speech in their Chinese learning. It was hoped that the data would help to explore the feasibility of using synthesized speech in the teaching Chinese as a foreign language.

2.1 Participants

In total, 42 students from the University of Notre Dame, Middle Tennessee State University, and Texas Tech University were recruited to participate in the study. After data collection, it turned out that 3 out of those 42 participants were native or bilingual heritage speakers. Their data were excluded in the final data analysis in order to maintain the homogeneity of the samples and to guarantee the reliability of the results.

Although the data was collected across three American universities, the participants' achieved similar beginning or intermediate level Chinese language proficiency due to similar course settings, requirements, and teaching pace. At the University of Notre Dame where *Integrated Chinese* is adopted, both the beginning and intermediate level Chinese classes meet five times per week. At Texas Tech University, the beginning level Chinese classes also meet five hours per week, whereas the intermediate level classes meet three times per week. Similar to Texas Tech University, classes at both the beginning and intermediate level at Middle Tennessee State University meet three hours per week. The latter two universities use the same textbook *Chinese Link* for instruction. Students from the University of Notre Dame and Texas Tech University were either majoring or minoring in Chinese, while all participants from the Middle Tennessee State University were non-major or non-minor students at the time when the experiment was conducted in Spring 2022.

Among the 39 participants that were included in the final data analysis, 17 students were recruited from the University of Notre Dame, 10 from Middle Tennessee State University, and 12 from Texas Tech University. Twenty-eight students were taking beginning level Chinese language classes at their respective universities, 10 were taking intermediate level Chinese language classes, and 1 student didn't report the Chinese language class he/she was taking. At the time when this study was conducted, the beginning level students had all completed one semester of classes, either 45 (3 hours/week) or 75 (5 hour/week) hours of class instruction, and the intermediate level students had completed three semesters, either 135 (3 hours/week) or 225 (5 hour/week) hours of class instruction. Although the course settings among the three universities were different, it is safe to consider all participants' language proficiency to have been at either the beginning or intermediate level. None of the participants were advanced-level learners.

In terms of previous experience using speech processing technologies in their native languages, 37 out of 39 (95%) participants had used smart devices such as iPhone, Google TV and/or Amazon Echo, etc. Thirty-six participants (92%) were aware of the voice-enabled functions in their native languages on the smart devices. Thirty-two (82%) participants were aware that some video clips on YouTube were dubbed with computer-generated speech in their native languages, and thirty-three participants (85%) had watched video clips or heard audio clips that were dubbed with computer-generated speech in their native languages. As compared with their experiences in using speech processing technologies in their native languages, a large portion of participants also had similar experiences in Chinese speech processing technologies. Twenty-six (64%) participants had watched video clips or heard audio recordings dubbed with computer-generated speech in Chinese. Twenty-five (64%) had used computer-generated speech in their Chinese learning. These data suggest that the majority of participants in this study have had rich experiences in using speech processing technologies in both their native languages and Chinese.

2.2 Materials, research design, and data collection

2.2.1 Materials

In this study, 10 audio clips were used as audio prompts to examine whether the participants could identify audio recordings made by humans and those generated by computers. Among them, 5 audio clips were recorded by humans and the other 5 were generated using Tencent's Text-to-Speech (TTS) technology[2]. The 5 audio clips recorded by real people were downloaded from *China Plus*[3], the official English news website of China Radio International, and *Chinese Take-In*[4], online learning materials for first-year Chinese learners developed by the University of Texas at Austin.

In order to minimize the effect of participants' familiarity with the materials on the identification task, 4 unfamiliar sentences beyond the participants' current language levels (as judged by the authors) were mixed together with the other 6 sentences that were deemed to be comprehensible to all participants. The transcripts of the 10 audio clips and their attributes are presented in Table 1 below.

Table 1 Transcripts of the ten audio clips used in the study

Transcripts of the audio clips	Attributes
Q6: 你是老师吗？	CG
Q7: 我不是中国人。	RP
Q8: 明天起冷空气东移南下驱散北方雾霾。	RP
Q9: 法德领导人就应对欧元区债务危机采取全面有力的解决方案达成一致。	CG
Q10: 我是中国人。	CG
Q11: 我是老师。	CG
Q12: 我也没有哥哥。我有一个弟弟。	RP
Q13: 明天起冷空气东移南下驱散北方雾霾。	CG
Q14: 法德领导人就应对欧元区债务危机采取全面有力的解决方案达成一致。	RP
Q15: 我是美国人。	RP

[2] Tencent's Text-to-Speech technology: https://cloud.tencent.com/product/tts_(in Chinese) or https://www.tencentcloud.com/products/tts (in English).

[3] China Plus: https://chinaplus.cri.cn/

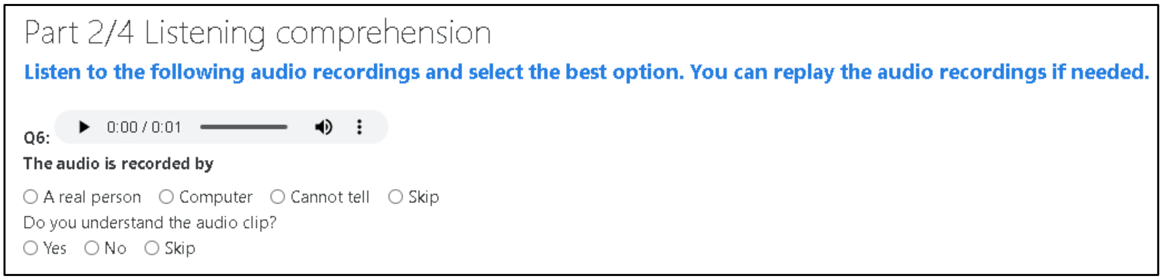
[4] Chinese Take-In: https://www.laits.utexas.edu/chinese_take_in/about.php

Note: 1) CG stands for computer-generated, and RP real person voice; 2) Q6, Q7, … represents the specific question number as they appeared the online survey (to be discussed below).

As can be seen in Table 1, both Q8 (read by real person) and Q13 (generated by computer) had the same content that were beyond the participants' current language proficiency levels, as were Q9 and Q14. All the other sentences would be comprehensible to all the participants since they were all covered in the beginning level Chinese textbooks used at the three American universities.

2.2.2 Experiment design and procedure

An online survey was used to answer the research questions regarding the feasibility of using synthesized speech in teaching Chinese as a foreign language, especially at the beginning and intermediate levels. The survey consisted of four parts. The first part collected participants' language background and CSL learning information, such as their current Chinese levels, textbooks they were using, and Chinese language courses they were taking, etc. The second part contained questions that asked participants to identify audio recordings made by a real person and computer (c.f., Figure 1) and self-report their comprehension. The third part asked participants to read aloud 12 lines of text and used Speech-to-Text technology to recognize their speech [5]. The fourth part investigated participants' experiences (with Yes/No questions) and opinions (using 5-point Likert scales) towards using speech processing technologies in both their native languages and Chinese (c.f., Table 2).


Part 2/4 Listening comprehension

Listen to the following audio recordings and select the best option. You can replay the audio recordings if needed.

Q6: ▶ 0:00 / 0:01

The audio is recorded by

○ A real person ○ Computer ○ Cannot tell ○ Skip

Do you understand the audio clip?

○ Yes ○ No ○ Skip

Figure 1 Part 2 of the online survey

The survey was developed and made available on a secured server at Middle Tennessee State University. Following the institutions' IRB policies, the participants from the three universities completed the survey anonymously either at home or in language labs based on their personal schedules and preferences. Before the study, all participants read and signed the informed consent form online. No personally identifiable information was collected throughout the process.

[5] The third part is not related directly to the research questions at hand and hence the results is not reported in this paper.

Table 2 Questions about participants’ previous experiences and attitudes

Experience in native languages	Q31. Are you aware that some video clips on YouTube are dubbed with computer-generated speech in your native language?
	Q32. Have you watched video clips such as on YouTube or heard audio clips that are dubbed with computer-generated speech in your native language?
Experience in Chinese	Q33. Have you watched video clips such as on YouTube or heard audio recordings dubbed with computer-generated speech in Chinese?
	Q34. Have you ever used computer-generated speech in your Chinese learning?
Attitude	Q35. Do you agree that computer-generated speech can help improve your Chinese listening skills?
	Q36. Would you object if your Chinese teacher used computer-generated listening materials in your Chinese class?

2.2.3 Variables and data analysis

Seven (7) variables were used for data analysis, including 1) the language proficiency levels of participants (labeled as “LEV” with 2 levels, beginning vs. intermediate), 2) the language difficulty of the audio clips (labeled as “DFT” with 2 levels, comprehensible at or beyond their current proficiency levels), 3) the frequency of participants’ correct identification of the audio clips (labeled as “JDG”, i.e., if they can tell if a particular audio clip was made by a human or computer), 4) participants’ self-reported comprehension of the 10 audio clips (labeled as “UDS”), 5) participants’ self-reported experiences of using synthesized speech in their native languages (labeled as “EPN”), 6) participants’ self-reported experiences of using synthesized speech in Chinese (labeled as “EPC), and 7) participants’ attitudes towards the use of synthesized speech (labeled as “ATT”) for Chinese language learning.

Regarding the language proficiency levels (LEV), beginning level was coded as “1” and intermediate level as “2”. As for the language difficulty (DFT), the 4 sentences which were deemed to be beyond participants’ comprehension were coded as “2” and the other 6 were coded as “1”. The numerical variables, identification (JDG), and comprehension (UDS) were collected from participants’ answers to questions in Part 2. If a correct identification of computer speech was made, 1 point was given. All correct points were added as the value of JDG. The self-reported understanding of the audio clips was coded as “1” and no-comprehension was coded as 0. The sum of all audio comprehension was the value of UDS. As long as a student reported he/she had experience in synthesized speech in his/her native languages (EPN), it was scored “1”. Thus the value of EPN (experiences) was the sum of the two questions for EPN (Q31 and Q32, c.f., Table 2), with 1 for each Yes answer. The same coding method was also applied to the value of the numerical variable EPC (Q33 and Q34). As to the 5-point Likert scale variable ATT, the combined mean of the two questions was calculated. All skipped answers were coded as missing data.

Three hypotheses were created for data analysis:

(1) Students can distinguish between audios recorded by a computer and a real person, and their judgement is not influenced by their Chinese proficiency levels, comprehension of the audio recordings, previous experiences on synthesized speech either in their native languages or Chinese, and their attitudes;
(2) language difficulty, whether within or beyond the participants’ current language levels, has no effect on participants’ correct identification of computer-generated speech; and
(3) under the circumstance that participants can correctly identify audio recordings made by a computer or a real person, most students will still show positive attitude towards the use of synthesized speech in Chinese language learning and teaching.

A Pearson correlation coefficient *r* was calculated to test hypothesis (1), i.e., the relationships of JDG with LEV, UDS, EPN, EPC, and ATT. An independent sample *t*-test was performed to test hypothesis (2). The percentage of the students’ attitudes was calculated to measure hypothesis (3).

2.3 Results

The Pearson correlation coefficients were computed to assess the relationship among participants’ language proficiency levels (LEV), their correct identification of the audio clips (JDG), their understanding (UDS) of the audio prompts, their previous experience in native languages (EPN), their previous experience in Chinese (EPC), and their attitudes towards synthesized speech. The results are shown in Table 3 below.

Table 3 Pearson correlation among JDG, UDS, STN, STC, and ATT

	JDG	LEV	UDS	EPN	EPC
LEV	-.12				
UDS	.18	.29			
EPN	.06	.37*	.05		
EPC	-.09	-.17	-.05	.06	
ATT	-.31	.21	-.13	.18	.01

* p< 0.05

There was a positive correlation between the students’ language levels and their previous experiences in their native languages, $r = 0.7$, $n = 37$, $p < 0.05$. i.e., participants with higher Chinese language proficiency levels also had more experience with synthesized speech in their native languages. No other positive or negative correlations were found. That is, participants’ language proficiency levels did not influence their judgement of the audio clips, whether the audios were recorded by human or computer. Their judgement was neither affected by their’ understanding of the audio clips nor their previous experiences with synthesized speech in their native languages or in Chinese. Further, participants’ attitudes towards using synthesized speech in teaching and learning Chinese did not influence their judgement of the audio clips.

Table 4 presents the rate of correct identification and participants' self-reported understanding of the audio clips. As can be seen from the table, the participants' identification of computer speech is not 100%. While participants' correct identification of all audio clips reached 64.6% (or 2 out of 3 clips), their successful identification of the four sentences that were deemed to be beyond their language levels was 56.4%, and 70.0% of the six sentences within their current language level. Further, only one out of five (or 21.1%) of the participants could understand the four sentences beyond their level, whereas 95.3% of the participants could understand the sentences within their language proficiency levels.

Table 4 Correct identification and understanding

Questions	Type	Level	Correctness %	Under-standing %
Q6: 你是老师吗？	computer	within	76.9	97.4
Q7: 我不是中国人。	person	within	74.4	97.4
Q8: 明天起冷空气东移南下驱散北方雾霾。	person	beyond	51.3	33.3
Q9: 法德领导人就应对欧元区债务危机采取全面有力的解决方案达成一致。	computer	beyond	69.2	12.8
Q10: 我是中国人。	computer	within	53.8	92.3
Q11: 我是老师。	computer	within	33.3	92.3
Q12: 我也没有哥哥。我有一个弟弟。	person	within	92.3	97.4
Q13: 明天起冷空气东移南下驱散北方雾霾。	computer	beyond	79.5	30.8
Q14: 法德领导人就应对欧元区债务危机采取全面有力的解决方案达成一致。	person	beyond	25.6	7.7
Q15: 我是美国人。	person	within	89.7	94.9
Average			**64.6**	**65.6**

An independent sample *t*-test was conducted to evaluate the hypothesis that no matter how difficult the sentences were, i.e., either within or beyond their current language proficiency levels, the frequency of participants' correct identification of the audio clips (made by human or computer) would be the same. The test was non-significant, t (8) = 0.92, p = 0.38. The result met the assumption that the rate of correct identification does not differ by the two levels of sentence difficulty, i.e., within or beyond their language proficiency levels. The rate of correct identification on the within-level sentences (M = 70.07, SD = 22.64) was the same as that on the beyond-level sentences (M = 56.40, SD = 23.61).

Another independent sample *t*-test was conducted to evaluate whether participants' self-reported understanding of the sentences matched well with the difficulty of the sentences. The test was significant, *t* (8) = 14.21, *p* = 0.00. The result showed that participants' understanding of the within-level sentences (*M* = 95.28, *SD* = 2.51) was significantly higher than that of the beyond-level level sentences (*M* = 21.15, *SD* = 12.80).

Table 5 Results of the sample t-tests on the variables of DFC, UDS, and JDG

Logistic parameter	Within level		Beyond level		*t* (8)	*p*	Cohen's *d*
	M	*SD*	*M*	*SD*			
JDG	70.07	22.63	56.40	23.61	0.92	.38	12.00
UDS	95.28	2.51	21.15	12.80	14.21	.00*	8.08

* p< 0.05

Two 5-point Likert scale questions were asked to assess the participants' attitudes towards using synthesized speech in Chinese language teaching and learning. The majority (79.5%) agreed (to various degrees) on the helpfulness of computer-generated speech for Chinese language teaching and learning (*M* = 3.36, *SD* = 1.09, *n* =39, c.f. Figure 2). When asked if their Chinese teachers could use synthesized speech as listening materials in Chinese classes, 74.3% of the participants did not object to the idea (*M* = 4.0, *SD* = 1.1, *n* = 39, c.f., Figure 3). These results indicate that the participants hold a positive view on the use of computer-generated speech in the Chinese classroom.

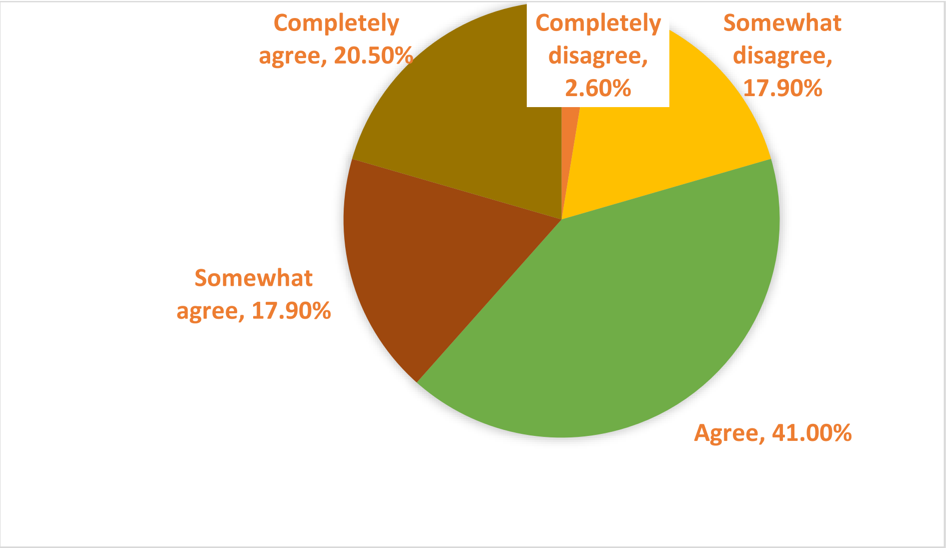


Figure 2 Responses to Q35: Do you agree that computer-generated speech can help improve your Chinese listening skills?

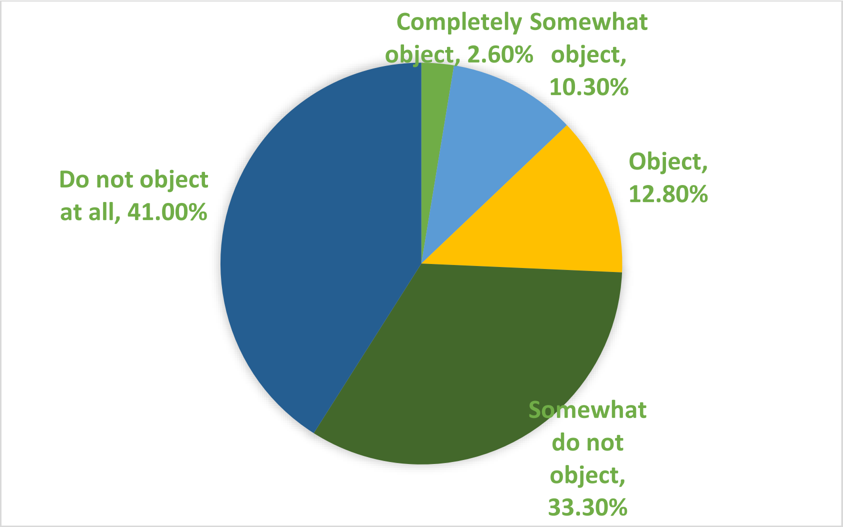


Figure 3 Responses to Q36: Would you object if your Chinese teacher used computer-generated listening materials in your Chinese class?

2.4 Discussions

It is not surprising that participants were able to correctly identify computer-generated speech most of the time, even though they were not 100% successful. Participants' failure to correctly identify certain sentences suggests that those synthesized audios are similar enough to human speech that they have heard before. Due to limited time and resources, this study did not ask participants to reflect on what criteria the participants used to make their judgement. Based on the authors' personal experiences, prosodic features (such as intonation, tempo of speech, and even pauses) that make speech natural are most likely the traits that participants used when making their judgement, whereas the pronunciation of individual sounds or syllables is less likely to betray the speaker of the audio recordings.

The fact that participants did not hold a negative view on the use of synthesized speech in their CSL learning is no surprise as well. Their experiences with the real-world use of TTS technology in their native languages may have paved the way for their acceptance of using the same technology for Chinese language learning. Even though this study did not ask in what ways synthesized speech could be helpful for their CSL learning, it is conceivable that the potential of TTS technology in providing additional materials to meet their learning needs is appealing enough.

3. Implications for SLA and CSL instruction

This study has found that beginning and intermediate level CSL learners can differentiate between audio recordings made by humans and computers. They can also comprehend synthesized Chinese speech that are within their current language proficiency levels. Further, most of the participants have experiences with speech processing technologies in both their native languages and Chinese, and welcome the use of TTS technology for their Chinese language learning. Findings from this study have several implications for SLA research and CSL classroom pedagogy.

To begin, given the accepting attitude of CSL learners, instructors can be assured to experiment with Text-to-Speech technology to produce listening materials to meet the needs of CSL learners. These listening materials could, for example, serve as additional or supplemental materials for CSL learners.

If an instructor decides to introduce synthesized speech into the classroom, he/she should be aware that there is not enough research on the best phase in time for synthesized speech. For example, is it better to use it as supplemental materials at higher levels of instruction rather than the beginning level?

In addition, given the fact that Text-to-Speech technology has been increasingly used for real world communication by native speakers, instructors should begin to consider the necessity and possibility of including comprehension of computer-generated speech as part of the learning outcomes, just like how standard pronunciation and dialects are treated in the language curriculum design.

Moreover, this study only used Tencent's speech technology to generate the audio prompts. While the authors decided that the speech quality is natural enough for both instruction and this study, instructors are encouraged to experiment with similar technologies from other vendors such as iFlyTEK, Baidu or Microsoft, etc. While the authors have found that the speech quality from those different vendors have reached more or less the same maturity level, these technologies may not share the same accessibility. Further, due to hardware and software availability, it is likely that instructors may be confined to one technology vendor even though in theory there are many options available.

Finally, instructors should be aware that even though Text-to-Speech technologies are available in many languages, and learners are likely to have encountered them in their native languages, there are still a series of second language acquisition questions that need to be studied thoroughly before full confidence can be given to the use of the technology in second language learning and instruction. For example, can synthesized speech be considered authentic? Can second language learners develop or improve their listening skills equally as well with the help of synthesized speech? Will robotic speech have a negative impact on learners' pronunciation and accent? What are the side effects of errors introduced in synthesized speech or deviations (such as intonation or pronunciation, etc.) from real human speech on learners' language development? And more generally, what are the possible positive or negative effects on learning outcomes when Text-to-Speech technology is fully integrated as part of the technical assistance for second language learning and instruction?

It will be through experimenting with the technology in second language classrooms that these questions and issues can be understood and addressed.

4. Conclusions

This study examined CSL learners' ability to differentiate between human speech and computer-generated speech, and their attitudes towards the use of synthesized speech for CSL learning and instruction. The data from this study suggests that while CSL learners, even at the beginning or intermediate levels, do have the ability to tell the difference between human and computer speech, they welcome the use of Text-to-Speech technology for their CSL learning. Future research is needed to overcome some limitations in this study and further explore the issue of applying Text-to-Speech technology for its use in second language education.

First, only TenCent technology was used to generate the audio materials. As discussed in Section 3, even though both the quality and naturalness of speech synthesized by other vendors such as Microsoft or Google are judged by the researchers of this study as comparable, it is still desirable to experiment with synthesized speech using other vendors' technology to see if CSL learners also find them equally intelligible, and whether the varied qualities of synthesized speech would affect their attitudes towards the use of the technology for their Chinese language learning.

Secondly, this study involved beginning and intermediate level CSL learners but not learners at more advanced levels. It is necessary to examine the latter group of learners to fully understand and determine if language proficiency would affect learners' perception of intelligibility and their attitudes towards Text-to-Speech technology.

Thirdly, participants in this study involved CSL learners only. The case with native speakers was not studied. It would be interesting to experiment with native speakers so that findings concerning the latter can serve as a reference point for fully understanding the issues in the application of TTS technology.

Finally, due to limited time and resources allowed in conducting the survey, this study only used a small number of audio clips (5 by human and computer, respectively). Hence any statistical claims made based on the small set of data should best be interpreted as representing a trend rather than a definite claim about the reality.

Despite these limitations, findings from this study confirm that it is viable to experiment with Text-to-Speech technology for Chinese language learning and instruction. Future research is needed to better understand the issues and circumstances involved in using speech technologies for second language education.

References

Bione, T., Grimshaw, J., & Cardoso, W. (2016). An evaluation of text-to-speech synthesizers in the foreign language classroom: learners' perceptions. In S. Papadima-Sophocleous, L. Bradley, & S. Thouësny (Eds.), *CALL communities and culture – short papers from EUROCALL 2016* (pp. 50-54). Research-publishing.net. https://doi.org/10.14705/rpnet.2016.eurocall2016.537

Cardoso, W., Smith, G., & Garcia Fuentes, C. (2015). Evaluating text-to-speech synthesizers. In F. Helm, L. Bradley, M. Guarda, & S. Thouësny (Eds.), *Critical CALL – Proceedings of the 2015 EUROCALL Conference, Padova, Italy* (pp. 108-113). Research-publishing.net. http://dx.doi.org/10.14705/rpnet.2015.000318

Gardner, R. C. (1968). Attitudes and motivation: Their role in second-language acquisition. *TESOL Quarterly, 2*(3), 141-150.

Huang, Y. C., & Liao, L., C. (2015). A study of Text-to-Speech (TTS) in children's English learning. *Teaching English with Technology*, 15(1), 14-30.

Kent, D. (2021). Voice-user interfaces for TESOL: Potential and receptiveness among native and non-native English speaking instructors. *Language Learning & Technology, 25*(3), 27–39. http://hdl.handle.net/10125/73444

Liakin, D., Cardoso, W., & Liakina, N. (2017). The pedagogical use of mobile speech synthesis (TTS): Focus on French liaison. *Computer Assisted Language Learning, 30*(3-4), 348–365. https://doi.org/10.1080/09588221.2017.1312463

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

Soon, G. Y., Warris, S. N., & Al Marimuthu, R. (2020). Chinese language learners' intrapersonal and interpersonal perceptions of a pinyin text-to-speech system. In M. R. Freiermuth & N. Zarrinabadi (Eds.), *Technology and the psychology of second language learners and users* (pp. 381–401). Palgrave Macmillan/Springer Nature. https://doi.org/10.1007/978-3-030-34212-8_15

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

Wikipedia. (n.d.). *Speech synthesis*. https://en.wikipedia.org/wiki/Speech_synthesis

Wood, S. G., Moxley, J. H., & Wagner, R. K. (2017). Does use of Text-to-Speech and related read-aloud tools improve reading comprehension for students with reading disabilities? A meta-analysis. *Journal of Learning Disabilities, 51*(1). https://journals.sagepub.com/doi/10.1177/0022219416688170

Yeh, R. (2014). Effective strategies for using Text-to-Speech, Speech-to-Text, and Machine-Translation technology for teaching Chinese: A multiple-case study (UMI Number: 3666758). [Doctoral dissertation, Northcentral University]. ProQuest Dissertations Publishing.