Evaluating the usage of Text-To-Speech in K12 education

Dai, Laduona; Kritskaia, Veronika; Velden, Evelien van der; Jung, Merel M.; Postma, Marie; Louwerse, Max M.

Published in:
ICEEL '22: Proceedings of the 2022 6th International Conference on Education and E-Learning

DOI:
10.1145/3578837.3578864

Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

Evaluating the usage of Text-To-Speech in K12 education

Laduona Dai
L.Dai_1@tilburguniversity.edu
Tilburg University
Netherlands

Veronika Kritskaia
V.Kritskaia@tilburguniversity.edu
Tilburg University
Netherlands

Evelien van der Velden
E.H.A.vdrVelden@tilburguniversity.edu
Tilburg University
Netherlands

Merel M. Jung
M.M.Jung@tilburguniversity.edu
Tilburg University
Netherlands

Marie Postma
Marie.Postma@tilburguniversity.edu
Tilburg University
Netherlands

Max M. Louwerse
m.m.louwerse@tilburguniversity.edu
Tilburg University
Netherlands

ABSTRACT

With increased interest in the use of virtual avatars for educational purposes, there is a growing need for high-quality text-to-speech solutions. However, the effects of using synthesized speech in educational applications for younger listeners are still unclear as past findings have been inconsistent and most of them have been obtained in a lab setting with adult assessors. Next to that, it is unclear how much training material is needed for high quality speech synthesis. Particularly for low resource languages, the assumption that good quality synthesized speech requires substantial amounts of vocal recordings to train may be hindering the development of TTS-based solutions. In this study, we created four Dutch text-to-speech (TTS) models from different amounts of training material and evaluated the models in terms of voice perception and recall with K12 students in a classroom environment. Results showed that while the original human voice outperformed the synthesized voices in terms of the listening experience and knowledge test score, more hours of training material did not necessarily result in better outcomes suggesting that 10-15 hours of speech material might be sufficient for training a Dutch TTS. A weak positive correlation was found between listening experience and knowledge test performance, with the low listening effort being the most important factor. This outcome suggests that comprehensibility is likely the most important TTS feature for educational applications.

CCS CONCEPTS

• **Applied computing** → **E-learning**; • **Social and professional topics** → **Children**.

KEYWORDS

Text-to-speech, K12 education

ACM Reference Format:
Laduona Dai, Veronika Kritskaia, Evelien van der Velden, Merel M. Jung, Marie Postma, and Max M. Louwerse. 2022. Evaluating the usage of Text-To-Speech in K12 education. In *2022 6th International Conference on Education and E-Learning (ICEEL 2022), November 21–23, 2022, Yamanashi, Japan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3578837.3578864

ICEEL 2022, November 21–23, 2022, Yamanashi, Japan

ACM ISBN 978-1-4503-9842-8/22/11.
https://doi.org/10.1145/3578837.3578864

1 INTRODUCTION

In the past, text-to-speech (TTS) systems have been used in education for many different purposes, including support for students with reading difficulties [2, 4, 38, 40] or as assistive technology within intelligent tutoring systems [8, 22, 30]. Despite many available studies on the usage of synthesized speech for multimedia applications, the findings on whether synthesized speech facilitates learning are still inconsistent. Some studies indicate that human speech outperforms synthesized speech. For example, [3, 24] reported that young and young adult learners achieved higher learning outcomes with a human voice. Similarly, the results from a meta-analysis [11] suggest that participants perform better in the near transfer of knowledge with pedagogical agents using real human voice compared to synthesized voice. In contrast, two meta-analyses of research on pedagogical agents ([31] and [6]) both revealed no significant differences in learning outcomes in conditions using human speech and generated speech. Finally, in a study by [8], participants learning with a modern TTS voice outperformed a classic TTS voice and human voice on knowledge transfer, suggesting that the quality of the TTS system plays an important role in the learning process.

Almost all past studies utilized available TTS voices from commercial software (e.g., Microsoft [8, 14], IVONA [20, 21] and NeoSpeech [8]) and most of them evaluated English voices only. Commercial software clearly has its limitations since it usually provides only a limited number of voices and few options for customization. Meanwhile, studies have shown that the characteristics of the voice could affect users' perception of the voice and potentially their willingness to use the application. For example, [34] observed that the gender of the synthesized voice may affect intelligibility. Interestingly, [36] also suggested that synthesized speech should be evaluated in the environment in which it is intended to be used because the noise in the environment is likely to impact the TTS perception. With respect to linguistic features of the system, [28] showed that using sociolects and dialects for synthesized speech applications could be beneficial. Adding personality to the voice, even an existing identity – for example, the voices of famous astronauts for learning applications about the space – can further enhance the learning experience. Needless to say, commercial options for customized TTS voices can be expensive and providing many hours of training material might pose privacy concerns. Training TTS voices with desired

speaker characteristics on local machines using state-of-the-art architectures might provide a solution to these issues. Furthermore, the ability to synthesize good quality speech with limited hours of training material is appealing for many low-resource languages where large audio corpora may not be available, thus making it difficult to replicate the English results. The development of neural TTS (i.e., speech generated with neural networks) in recent years has made it possible to produce high-quality synthesized speech using a reasonable amount of training material with models like Google Tacotron 2 [33] and VITS [16]. Recently, speech synthesized by the VITS architecture trained on a 24-hour single speaker English dataset was evaluated as having near-human quality [16]. Studies in this field often focus on improving the Mean Opinion Score (MOS) [19], a common method to evaluate the quality of TTS systems. However, it is also important to evaluate the capability of these modern TTS models not only in a controlled lab environment, but also in real-world scenarios with target users (e.g., students in a noisy classroom) and specific tasks (e.g., recall of information offered by the TTS system in a knowledge test). This study aimed to evaluate the effects of four Dutch TTS models trained from different amounts of speech material in comparison to the human voice in the classroom environment.

2 BACKGROUND

2.1 Perception of synthetic voice by listeners

Can TTS system ever replace natural human voice? Past studies suggest the human brain reacts differently to synthetic voice and natural voice. For example, recall performance for a list of ordered words was found to be lower for synthetic speech compared to natural speech and synthetic speech was less accurately identified [37]. It seems that encoding difficulties are the main reason for reduced memory performance, supporting the idea that more processing capacity is needed to code synthetic speech than natural speech. Findings of a study with 8–11 year old children suggest that the human brain might prioritize the processing of human voice signals over speech sounds in general, and that natural speech might be more efficiently processed than synthetic speech [39].

On the other hand, results of one of the early studies by [32], showed that participants who received training by a synthetic voice outperformed those who were trained with a human voice and those that received no training in a post-test task of word recognition and showed improved long-term knowledge retention after six months. In a more recent study, high school students in Japan showed a higher retention rate when using TTS as an English reading aloud tool compared to no instruction groups [15]. Moreover, a TTS voice was found to be as effective as a human voice for short-term and long-term factual medical knowledge learning and the use of TTS instead of a human narrator saved 48,000 USD [26]. English teaching in the classroom with TTS significantly improved word stress, word intonation, pitch contour, and fluency of reading [25]. No significant differences were found between students who learned with a TTS voice compared to a human voice in terms of learning outcome [9], and results in [35] suggested participants could acquire Dutch pronunciation on a short-term basis using Google Translate's TTS function.

Conflicting findings can potentially be explained by the effect of TTS quality. One of the most famous TTS architectures in recent years is Tacotron 2, published in 2018 [33]. With 24.6 hours of training material from a single speaker non-public English dataset and the usage of 32 GPUs, the model achieved a MOS of 4.53 (out of 5) compared to the human voice (ground truth) of 4.58. A more recent study showed better TTS quality could be achieved with even less computational power. The end-to-end TTS architecture VITS [16], published in 2021, achieved a MOS of 4.43 compared to Tacotron 2 with a MOS of 4.25 and a ground truth of 4.46 when using a 24-hour publicly available single speaker English dataset trained with 4 GPUs. Results from studies that compare off-the-shelf TTS systems also provide some interesting insights. [1] in 2019 evaluated the performance of eight available TTS systems from Google, Microsoft, Ivona, Loquendo, Espeak, Pico, AT&T, and Nuance for the use in robots to check the voice intelligibility, expressiveness, artificiality, and suitability. The results from 125 participants suggest that the voice from Google's TTS outperformed other systems on intelligibility and expressiveness, and it was also rated as the least artificial one. In 2020, [5] compared 18 TTS voices in terms of quality using long-form content with more than one thousand participants. The TTS voices include off-the-shelf ones from Amazon Polly, IOS, Microsoft, and Google Cloud, and synthetic voices trained from Tacotron and Tacotron 2. Evaluation results suggest the best MOS-rated synthetic voices are trained from published TTS models and rated higher than the best off-the-shelf voice from Google Cloud. The analysis also revealed that high-quality TTS voices are close to human voices and in some situations may be more preferable.

Another factor that might have played a role in previous studies is the effect of age. [27] indicated the different patterns of intelligibility tasks among 7–8 year olds, 11–12 year olds, and adults. The effects of age differences were also found in terms of response latencies by another study when using synthetic speech. [29] investigate how children from 6–7 and 9–11 year olds comprehend speech by comparing speech response latencies with sentence verification tasks after some listening practice at two listening times spaced between 3 and 7 days apart. For the synthetic speech condition, only the younger children group showed response latencies shortened significantly from time 1 to time 2. Apart from the studies conducted in the 90s, relatively more recent studies also discovered the differences when using TTS for different age groups. For example, [12] evaluated the intelligibility of speech in single words and sentences under background noise conditions for children 3-5 years old. Results suggested that 4- and 5-year-old children perform better than 3-year-old children, suggesting that TTS perception might improve with age.

2.2 Contextual appropriateness and voice characteristics

The characteristics of synthetic voice should match its context and application to maximize its functionality and potential. In the study by [36], researchers reviewed recent TTS evaluations and gave some suggestions for the research in this field. In particular, contextual appropriateness should be taken into account when considering the quality of synthetic speech since different environments and

use cases require different degrees of attention (e.g., silent vs. noisy environment or news reading style vs. conversation style). For the purpose of language learning, the generated voices should be clear with natural pronunciation and intonation. Additional voice characteristics could directly or indirectly affect learning when using synthetic speech. [17] showed that preschool children benefit more from an expressive voice that has more intonation and emotion than a flat voice when learning from a robot. [13] investigated how the length of breath sound affects recollection with synthetic speech. The authors used a modern concatenative TTS system comparing the respiratory sounds with lengths of 0ms, 300ms, and 600ms. Results revealed that the 600ms condition improved recollection and recollection was also better with shorter sentences than longer sentences.

2.3 Current study

The current study aims to evaluate the perception and learning effects of using synthesized Dutch voices generated from the VITS architecture [16] for K12 education in a classroom environment. In addition, we aim to answer the question of how much training material is needed to achieve acceptable performance in perception and learning. Specifically, the current study addresses three research questions: 1) How is synthesized Dutch voice generated from the VITS architecture evaluated by K12 students? 2) How does the number of hours of TTS training material affect students' perception of the voice? 3) Is there a relation between the perception of the TTS voice and recall?

3 METHOD

3.1 Participants

This study was approved by the Research Ethics and Data Management Committee within the university. Students' caregivers gave informed consent before the study and all students agreed to participate in the experiment. Two classes of in total 44 children aged 11–13 years (20 female, 24 male) from two Dutch primary schools were recruited to participate in the study. The participants had a mean age of 11.73 (SD = 0.54). As the entire school curriculum is in Dutch it is assumed that all students had a sufficient comprehension of the language.

3.2 Design

The within-subject experiment was conducted with all participants simultaneously. Each participant evaluated four TTS voices and the original human voice. The independent variable was *length of training material* with four levels corresponding to 5, 10, 15, and 20 hours. The dependent variables were the MOS [19] for evaluating speech perception and recall measured with the help of a knowledge test.

3.3 Materials

Voice fragments. The VITS architecture as implemented by the open speech technology company Coqui [1] was used for Dutch speech synthesis. In total, five different recordings were compared: a human voice from a Dutch audiobook of a historical novel and four TTS voices trained on 5, 10, 15, and 20 hours of material from the same audiobook (in the following sections, the four TTS systems will be denoted as TTS 5, TTS 10, TTS 15 and TTS 20). Each TTS voice generated two out of four selected fragments. The four fragments were different parts of learning material from a Dutch primary education publisher aimed at teaching K12 students about history. One fragment from the Dutch audiobook was selected as the human voice gold standard. All fragments were carefully selected to have a similar word length (78-98) and duration of the audio clip (25-30 seconds). In total, nine audio fragments were used: 4 TTS * 2 fragments + 1 fragment of the Dutch audiobook. All four TTS models were trained with the same training configurations, the only difference was the number of hours of training material. In a pilot experiment with a small sample of 10-12 year old children, participants indicated that the speech rate was too high to understand and remember the spoken content. Therefore, all synthesized and human speech was slowed down at 0.7 times to suit K12 students, in line with [23], who proposed that the mean speech rate for 11-year-old children should be 101.3 (SD=33.2) words per minute.

Mean opinion score. MOS is commonly used to evaluate the quality of TTS systems and consists of 7 items [19]. It is scored using a 5-point Likert-type scale designed to measure participants' perceptions of voice in terms of global impression, listening effort, comprehension problems, speech sound articulation, pronunciation, speaking rate, and voice pleasantness. All voice qualities are positively phrased so that higher scores indicate a more positive evaluation (see Figure 3 in Appendix). To help students to better understand and answer the questions, the original MOS questionnaire was translated into Dutch and the language was simplified.

Recall. Recall was assessed by knowledge questions related to the speech content of each fragment. Students answered one open-ended question and three fill-in-blank questions after listening to each of the five fragments. The answer format was either a single short sentence for the open questions or filling in one or two blanks. All answers were evaluated using the Damerau Levenshtein Distance score [10, 18]. The score calculated the similarity between the given responses and the correct answers on a scale from 0-100 (100 indicating a perfect match).

3.4 Procedure

Children were first informed about the purpose of the experiment, followed by the instructions, and signing of the consent form. Each participant received a printed version of the MOS questionnaire and knowledge test. Each audio fragment was played twice using a speaker placed in the classroom. After that, students were given enough time to complete the MOS and knowledge test questions. The entire experiment lasted about 20 minutes.

4 RESULTS

4.1 Mean Opinion Score (MOS)

The MOS was calculated by averaging the rating on all items as Cronbach's alpha for the different MOS sub-scales showed high internal consistency ($\alpha > .8$). Table 1 shows the MOS and standard

[1] https://github.com/coqui-ai/TTS

deviation for all four TTS models and the human voice evaluated by all students. The TTS model trained from 5 hours of materials had the lowest MOS , and the three TTS models trained from 10-hours, 15-hours, and 20-hours have a similar MOS. The human voice had the highest score of all models (see Figure 1).

Table 1: MOS per model with mode and standard deviation values.

Model	MOS	SD
TTS 5	2.299	1.162
TTS 10	2.721	1.024
TTS 15	2.786	1.197
TTS 20	2.769	1.119
Human voice	3.143	1.189

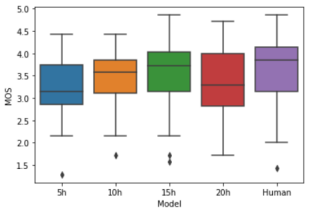

Figure 1: Box-plot for different models' MOS.

Table 2 shows the details of all seven MOS sub-scales for the different models. The human voice has the highest score on all voice characteristics. Among the different TTS models, TTS 10 is rated highest on listening effort and speech rate, TTS 15 on global impression, speech sound articulation, and pronunciation and TTS 20 on comprehension and voice pleasantness.

A Shapiro-Wilk test showed a non-normal distribution for TTS 5 model ($p < .05$). Thus, to compare MOS between different models and human voice, the non-parametric Friedman test was conducted. A chi-squared test showed a statistically significant difference in MOS between the four TTS models and the human voice ($\chi^2(4) = 34.72, p < .001$), and Kendall's W was 0.197 indicating a small effect size (degree of difference) according to Cohen's interpretation guidelines.

The Conover post-hoc test [7] was conducted for pairwise comparisons to determine significant differences between models. The false discovery rate method was used to adjust the p-values for multiple hypothesis testing at a 5% cut-off. The multiple pairwise comparisons showed statistically significant differences in MOS between TTS 5 and all other models and between the human voice and all TTS models (see Table 3). No statistically significant differences in MOS were found between TTS 10, TTS 15, and TTS 20.

4.2 Recall

Table 4 shows the average knowledge test scores and standard deviations for all four TTS models and the human voice. The human voice yielded the highest scores among all models (see Figure 2).

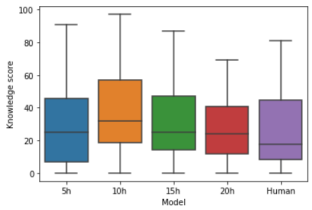

Figure 2: Box-plot of different models' knowledge test scores.

The TTS 10 model resulted in the lowest knowledge test score whereas TTS 5, TTS 15, and TTS 20 resulted in similar test scores.

A Shapiro-Wilk test showed a non-normal distribution for the knowledge scores of the TTS 10, TTS 20 models and the human voice ($p < .05$). Therefore, test scores between different models and the human voice were compared using the non-parametric Friedman test. A chi-squared test showed a statistically significant difference in knowledge test scores between the four TTS models and the human voice ($\chi^2(4) = 42.83, p < .001$), and Kendall's W was 0.243 indicating a small effect size (degree of difference) according to Cohen's interpretation guidelines.

As with the perception scores, the Conover post-hoc test was conducted for pairwise comparisons to determine which models are significantly different. The false discovery rate method was used to adjust the p-values for multiple hypothesis testing at a 5% cut-off. The multiple pairwise comparisons showed statistically significant differences between the human voice and all TTS models (see Table 5). There were no statistically significant differences found between the four TTS models.

Spearman's rank correlation was computed to investigate the relation between MOS and its sub-scores and the knowledge test performance. A weakly positive correlation was found between the overall MOS and the knowledge score ($r(218) = 0.219, p = .001$) (see Table 6). All MOS sub-scores were positively correlated with test scores. However, statistically significant correlations were only found for listening effort ($r(218) = 0.315, p < .001$), comprehension ($r(218) = 0.248, p < .001$), speech sound articulation ($r(218) = 0.252, p < .001$) and speech rate ($r(218) = 0.201, p = .003$).

5 DISCUSSION

Most previous TTS research efforts have focused on developing new neural network architectures to increase MOS ratings. However, most studies trained their systems in English only and evaluated the TTS with adult individuals in a lab setting. The current evaluation study was conducted with primary school students in classroom environments using four non-English TTS systems trained on different amounts of training material. The perception of the Dutch TTS models and performance on a knowledge test were compared to a human voice to investigate how different amounts of training material affect TTS perception and how voice perception affects recall.

The human voice used in our study yielded a MOS ($MOS = 3.143$) that was considerably lower than what has been reported in the literature for adult listeners (e.g., $MOS > 4.0$ in [16, 33]). This result suggests that vocal characteristics play an important role for younger listeners. Despite the low evaluation, the human voice

Table 2: Comparison of voice characteristics for the different TTS models. Bold numbers represent the highest ratings among the four TTS models. * indicates significant differences (i.e., $p < .05$) between human voice and the highest TTS score.

Voice characteristic	TTS 5	TTS 10	TTS 15	TTS 20	Human voice
Global impression	1.864	2.364	**2.386**	2.273	2.455
Listening effort	2.000	**2.750**	2.727	2.682	3.386*
Comprehension problems	2.659	3.023	3.205	**3.227**	3.568
Speech sound articulation	2.159	2.659	**2.932**	2.727	3.273
Pronunciation	2.136	2.523	**2.727**	2.705	2.841
Speech rate	3.614	**3.750**	3.455	3.568	4.068
Voice pleasantness	1.659	1.977	2.068	**2.205**	2.409

Table 3: Conover post-hoc analysis of MOS for all model pairs. * indicates that coefficients are significant ($p < .05$).

Model	TTS 5	TTS 10	TTS 15	TTS 20	Human voice
5h	-	0.008*	0.004*	0.017*	2.925e-07*
10h	-	-	0.703	0.759	8.280e-03*
15h	-	-	-	0.541	2.319e-02*
20h	-	-	-	-	5.213e-03*
Human voice	-	-	-	-	-

Table 4: Average knowledge test score and standard deviation for each TTS model and the human voice.

Model	Average Score (%)	SD
TTS 5	39.752	23.859
TTS 10	30.659	24.655
TTS 15	38.906	27.461
TTS 20	38.219	29.227
Human voice	66.131	22.107

scored better on all voice characteristics of the MOS sub-scales compared to the TTS models. However, most differences between the MOS sub-scale ratings for the human voice and the best TTS model were not significant. Only the listening effort was rated significantly worse for the TTS models. Moreover, TTS 5 – the TTS system trained on the least amount of material – had yielded the lowest scores among the TTS models for all voice characteristics. Interestingly, there were no significant differences between TTS 10, TTS 15, and TTS 20. These findings suggest that TTS systems trained on ten hours of material can achieve a similar voice quality to TTS systems trained on twenty hours of material.

As for the recall, the knowledge test scores for all four TTS models were relatively low compared to the human voice. No significant differences were found between the knowledge test scores of the different TTS models. Overall MOS ratings were weakly positively correlated with the knowledge test score. Four MOS sub-scale ratings correlated positively to knowledge test scores (listening effort, comprehension, speech sound articulation, and speech rate). The listening effort showed the highest correlation, indicating that low listening effort can lead to better recall. Considering that this vocal feature was the only one of the seven MOS sub-scales with a significant difference between the human voice and the highest rated TTS model, the listening effort seems to be an essential component of the MOS scale and a predictor for knowledge retention.

6 CONCLUSION

TTS research is a rapidly developing field and many studies have shown that state-of-the-art English TTS systems can achieve near human-level performance. In this study, we evaluated four Dutch TTS systems trained from different amounts of training material using a recent TTS architecture (VITS). We compared their evaluation to a human voice with K12 students in a classroom environment. The characteristics of the human voice were rated more positively compared to the synthesized voices even though the human voice received relatively low ratings compared to previous studies. The comparison of different TTS models indicated that more hours of training material do not necessarily result in better evaluations or increased learning performance suggesting that 10-15 hours of training material might be sufficient for training a Dutch TTS. The knowledge test results showed that students retained more information when listening to the human voice than to the TTS voices and a weak positive correlation was found between a positive listening experience and knowledge test performance. Lastly, reducing listening effort was found to be the most important factor to consider when using TTS voices for educating K12 students. These findings help develop TTS in primary school environments and serve as a reference for further improvements to customizable TTS systems.

ACKNOWLEDGMENTS

This research is part of the MasterMinds project, funded by the RegionDeal Mid- and West-Brabant, and is co-funded by the Ministry of Economic Affairs and Municipality of Tilburg. WPG Zwijsen provided the text materials for testing the TTS voices.

REFERENCES

[1] Fernando Alonso Martin, María Malfaz, Álvaro Castro-González, José Carlos Castillo, and Miguel Ángel Salichs. 2020. Four-features evaluation of text to speech systems for three Social Robots. *Electronics* 9, 2 (2020), 267. https://doi.org/10.3390/electronics9020267

[2] Saeed S. Alqahtani. 2021. Ipad text-to-speech and repeated reading to improve reading comprehension for students with SLD. *Reading & Writing Quarterly* (2021), 1–15. https://doi.org/10.1080/10573569.2021.1987363

[3] Robert K. Atkinson, Richard E. Mayer, and Mary Margaret Merrill. 2005. Fostering Social Agency in multimedia learning: Examining the impact of an animated agent's Voice. *Contemporary Educational Psychology* 30, 1 (2005), 117–139. https://doi.org/10.1016/j.cedpsych.2004.07.001

[4] Paola Bonifacci, Elisa Colombini, Michele Marzocchi, Valentina Tobia, and Lorenzo Desideri. 2021. Text-to-speech applications to reduce mind wandering in students with dyslexia. *Journal of Computer Assisted Learning* 38, 2 (2021), 440–454. https://doi.org/10.1111/jcal.12624

Table 5: Conover post-hoc analysis of knowledge score for all model pairs. * indicates that coefficients are significant ($p < .05$).

Model	TTS 5	TTS 10	TTS 15	TTS 20	Human voice
5h	-	1.127e-01	0.682	1.916e-01	4.271e-04*
10h	-	-	0.227	7.365e-01	3.714e-07*
15h	-	-	-	3.528e-01	7.804e-05*
20h	-	-	-	-	9.605e-07*
Human voice	-	-	-	-	-

Table 6: Spearman's correlation analysis between MOS and knowledge score. * indicates that coefficients are significant ($p < .05$).

Relation	Correlation Coefficient	p-value
MOS - Knowledge score	0.219	.001*
MOS (Global impression) - Knowledge score	0.002	.977
MOS (Listening effort) - Knowledge score	0.315	<.001*
MOS (Comprehension) - Knowledge score	0.248	<.001*
MOS (Speech sound articulation) - Knowledge score	0.252	<.001*
MOS (Pronunciation) - Knowledge score	0.025	.716
MOS (Speech rate) - Knowledge score	0.201	.003*
MOS (Voice pleasantness) - Knowledge score	0.067	.325

[5] Julia Cambre, Jessica Colnago, Jim Maddock, Janice Tsai, and Jofish Kaye. 2020. Choice of voices: A large-scale evaluation of text-to-speech voice quality for long-form content. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13. https://doi.org/10.1145/3313831.3376789

[6] Juan C. Castro-Alonso, Rachel M. Wong, Olusola O. Adesope, and Fred Paas. 2021. Effectiveness of multimedia pedagogical agents predicted by diverse theories: A meta-analysis. *Educational Psychology Review* 33, 3 (2021), 989–1015. https://doi.org/10.1007/s10648-020-09587-1

[7] William J. Conover. 1999. *Practical nonparametric statistics*. Wiley.

[8] Scotty D. Craig and Noah L. Schroeder. 2017. Reconsidering the voice effect when learning from a virtual human. *Computers and Education* 114 (2017), 193–205. https://doi.org/10.1016/j.compedu.2017.07.003

[9] Scotty D. Craig and Noah L. Schroeder. 2018. Text-to-speech software and learning: Investigating the relevancy of the voice effect. *Journal of Educational Computing Research* 57, 6 (2018), 1534–1548. https://doi.org/10.1177/0735633118802877

[10] Fred J Damerau. 1964. A technique for computer detection and correction of spelling errors. *Commun. ACM* 7, 3 (1964), 171–176.

[11] Robert O. Davis. 2018. The impact of pedagogical agent gesturing in Multimedia Learning Environments: A meta-analysis. *Educational Research Review* 24 (2018), 193–209. https://doi.org/10.1016/j.edurev.2018.05.002

[12] Kathryn D. Drager, Elizabeth A. Clark-Serpentine, Kate E. Johnson, and Jennifer L. Roeser. 2006. Accuracy of repetition of digitized and synthesized speech for young children in background noise. *American Journal of Speech-Language Pathology* 15, 2 (2006), 155–164. https://doi.org/10.1044/1058-0360(2006/015)

[13] Mikey Elmers, Raphael Werner, Beeke Muhlack, Bernd Möbius, and Jürgen Trouvain. 2021. Take a breath: Respiratory sounds improve recollection in synthetic speech. *Interspeech 2021* (2021). https://doi.org/10.21437/interspeech.2021-1496

[14] Zeng-Wei Hong, Yen-Lin Chen, and Chien-Ho Lan. 2014. A courseware to script animated pedagogical agents in instructional material for elementary students in English education. *Computer Assisted Language Learning* 27, 5 (2014), 379–394. https://doi.org/10.1080/09588221.2012.733712

[15] Harumi Kataoka, Makiko Ito, and Shigeru Yamane. 2015. Retention of English sentences learned by reading aloud using text-to-speech (TTS) speech sounds: A longitudinal study in a Japanese high school. *International Journal of Research Studies in Educational Technology* 5, 1 (2015). https://doi.org/10.5861/ijrset.2015.1331

[16] Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*. PMLR, 5530–5540.

[17] Jacqueline M. Kory Westlund, Sooyeon Jeong, Hae W. Park, Samuel Ronfard, Aradhana Adhikari, Paul L. Harris, David DeSteno, and Cynthia L. Breazeal. 2017. Flat vs. Expressive storytelling: Young children's learning and retention of a social robot's narrative. *Frontiers in Human Neuroscience* 11 (2017). https://doi.org/10.3389/fnhum.2017.00295

[18] Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, Vol. 10. Soviet Union, 707–710.

[19] James R Lewis. 2001. Psychometric properties of the mean opinion scale. *Proceedings of HCI International 2001: Usability Evaluation and Interface Design* (2001), 149–153.

[20] Tze Liew, Nor Azan Mat Zin, Noraidah Sahari, and Su-Mae Tan. 2016. The effects of a pedagogical agent's smiling expression on the learner's emotions and motivation in a virtual learning environment. *International Review of Research in Open and Distance Learning* 17, 5 (2016), 248–266. https://doi.org/10.19173/irrodl.v17i5.2350

[21] Tze Wei Liew, Su-Mae Tan, and Chandrika Jayothisa. 2013. The effects of peer-like and expert-like pedagogical agents on learners' agent perceptions, task-related attitudes, and learning achievement. *Educational Technology and Society* 16, 4 (2013), 275–286.

[22] Guido Makransky, Philip Wismer, and Richard E. Mayer. 2019. A gender matching effect in learning with pedagogical agents in an immersive virtual reality science simulation. *Journal of Computer Assisted Learning* 35, 3 (2019), 349–358. https://doi.org/10.1111/jcal.12335

[23] Isabel Pavão Martins, Rosário Vieira, Clara Loureiro, and M. Emilia Santos. 2007. Speech rate and fluency in children and adolescents. *Child Neuropsychology* 13, 4 (2007), 319–332. https://doi.org/10.1080/09297040600837370

[24] Richard E. Mayer. 2014. Principles based on social cues in multimedia learning: Personalization, voice, image, and embodiment principles. *The Cambridge Handbook of Multimedia Learning* (2014), 345–368. https://doi.org/10.1017/cbo9781139547369.017

[25] Hussein Meihami and Fateme Husseini. 2014. Bringing TTS software into the classroom: the effect of using text to speech software in teaching reading features. *Teaching English with Technology* 14, 1 (2014), 23–34.

[26] Stefan Minder, Michele Notari, Felix Schmitz, Rainer Hofer, and Ulrich Woermann. 2012. Computer Generated Voice-Over in a Medical E-Learning Application: The Impact on Factual Learning Outcome. *J. Univers. Comput. Sci.* 18, 3 (2012), 314–326.

[27] Pat Mirenda and David Beukelman. 1990. A comparison of intelligibility among natural speech and seven speech synthesizers with listeners from three age groups. *Augmentative and Alternative Communication* 6, 1 (1990), 61–68. https://doi.org/10.1080/07434619012331275324

[28] Michael Pucher, Gudrun Schuchmann, and Peter Fröhlich. 2009. Regionalized text-to-speech systems: Persona Design and Application scenarios. *Multimodal Signals: Cognitive and Algorithmic Issues* (2009), 216–222. https://doi.org/10.1007/978-3-642-00525-1_21

[29] Mary Reynolds and Lisa Jefferson. 1999. Natural and synthetic speech comprehension: Comparison of children from two age groups. *Augmentative and Alternative Communication* 15, 3 (1999), 174–182. https://doi.org/10.1080/07434619912331278705

[30] Susanne Schmidt, Gerd Bruder, and Frank Steinicke. 2019. Effects of virtual agent and object representation on experiencing exhibited artifacts. *Computers and Graphics* 83 (2019), 1–10. https://doi.org/10.1016/j.cag.2019.06.002

[31] Noah L. Schroeder, Olusola O. Adesope, and Rachel Barouch Gilbert. 2013. How effective are pedagogical agents for learning? A meta-analytic review. *Journal of Educational Computing Research* 49, 1 (2013), 1–39. https://doi.org/10.2190/ec.49.1.a

[32] Eileen C. Schwab, Howard C. Nusbaum, and David B. Pisoni. 1985. Some effects of training on the perception of synthetic speech. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 27, 4 (1985), 395–408. https://doi.org/10.1177/001872088502700404

[33] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, and et al. 2018. Natural TTS synthesis by conditioning wavenet on Mel Spectrogram predictions. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2018). https://doi.org/10.1109/icassp.2018.8461368

[34] Catherine Stevens, Nicole Lees, Julie Vonwiller, and Denis Burnham. 2005. On-line experimental methods to evaluate text-to-speech (TTS) synthesis: Effects of voice gender and signal quality on intelligibility, naturalness and preference. *Computer Speech & Language* 19, 2 (2005), 129–146. https://doi.org/10.1016/j.csl.2004.03.003

[35] Catharina van Lieshout and Walcir Cardoso. 2022. Google Translate as a tool for self-directed language learning. *Language Learning & Technology* 26, 1 (2022), 1–19. https://doi.org/10125/73460

[36] Petra Wagner, Jonas Beskow, Simon Betz, Jens Edlund, Joakim Gustafson, Gustav Eje Henter, Sébastien Le Maguer, Zofia Malisz, Éva Székely, Christina Tånnander, and et al. 2019. Speech synthesis evaluation — state-of-the-art assessment and suggestion for A novel research program. *10th ISCA Workshop on Speech Synthesis (SSW 10)* (2019). https://doi.org/10.21437/ssw.2019-19

[37] John A. Waterworth. 1985. Why is synthetic speech harder to remember than natural speech? *ACM SIGCHI Bulletin* 16, 4 (1985), 201–206. https://doi.org/10.1145/1165385.317493

[38] D. Heather White and Lorayne Robertson. 2015. Implementing Assistive Technologies: A Study on co-learning in the Canadian Elementary School Context. *Computers in Human Behavior* 51 (2015), 1268–1275. https://doi.org/10.1016/j.chb.2014.12.003

[39] Allison Whitten, Alexandra P. Key, Antje S. Mefferd, and James W. Bodfish. 2020. Auditory event-related potentials index faster processing of natural speech but not synthetic speech over Nonspeech analogs in children. *Brain and Language* 207 (2020), 104825. https://doi.org/10.1016/j.bandl.2020.104825

[40] Sarah G. Wood, Jerad H. Moxley, Elizabeth L. Tighe, and Richard K. Wagner. 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 (2017), 73–84. https://doi.org/10.1177/0022219416688170

A APPENDIX

Item	Content	1	2	3	4	5
1	*Global Impression: Your answer must indicate how you rate the sound quality of the voice you have heard.*	Bad	Poor	Good	Fair	Excellent
2	*Listening Effort: Your answer must indicate the degree of effort you had to make to understand the message.*	Message not understood with any feasible effort	Major effort required	Effort required	Slight effort required	No effort required
3	*Comprehension Problems: Your answer must indicate if you found single words hard to understand.*	Every word	Many	Some	Few	None
4	*Speech Sound Articulation: Your answer must indicate if the speech sounds are clearly distinguishable.*	No, not at all	No, not very clear	Fairly clear	Yes, clearly enough	Yes, very clearly
5	*Pronunciation: Your answer must indicate if you noticed any anomalies in the naturalness of sentence pronunciation.*	Yes, very annoying	Yes, annoying	Yes, slightly annoying	Yes, but not annoying	No
6	*Speaking Rate: Your answer must indicate if you found the speed of delivery of the message appropriate.*	No, too fast	No, too slow	Yes, but faster than preferred	Yes, but slower than preferred	Yes
7	*Voice Pleasantness: Your answer must indicate if you found the voice you have heard pleasant.*	Very unpleasant	Unpleasant	Fair	Pleasant	Very pleasant

Figure 3: MOS questionnaire