An evaluation of TTS as a pedagogical tool for pronunciation instruction: the 'foreign' language context

Tiago Bione[1], Jennica Grimshaw[2], and Walcir Cardoso[3]

Abstract. Despite positive evidence demonstrating the pedagogical benefits of Text-To-Speech (TTS) synthesisers for second/foreign language learning (Liakin, Cardoso, & Liakina, 2017), there is a need for up-to-date formal evaluations, specifically regarding its potential to promote learning. This study evaluates the voice quality of a TTS system in comparison with a human voice, and examines its pedagogical potential for use in an English as a Foreign Language (EFL) setting in terms of speech quality, ability to be understood by L2 users, and potential to focus on a specific language form. EFL learners in Brazil completed four tasks to evaluate the quality of TTS-generated texts. Results suggest that the TTS voice performed equally as well as the human voice in almost every assessment measure, demonstrating a high level of intelligibility and the ability to provide learners with opportunities to notice linguistic forms.

Keywords: text-to-speech synthesis, TTS, pronunciation, English as a foreign language.

1. Introduction

Second language (L2) researchers and practitioners have explored the pedagogical capabilities of TTS synthesisers – a type of speech technology that creates a spoken version of any written text – for their potential to enhance the acquisition of writing (Kirstein, 2006), vocabulary, reading (Proctor, Dalton, & Grisham, 2007), and pronunciation (Liakin et al., 2017). Despite positive evidence demonstrating the

1. Concordia University, Montréal, Canada; tiagobione@gmail.com
2. Concordia University, Montréal, Canada; jennica.grimshaw@gmail.com
3. Concordia University, Montréal, Canada; walcir.cardoso@concordia.ca

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

pedagogical benefits of TTS, there is a need for up-to-date formal evaluations, specifically regarding the potential for TTS to promote the conditions under which languages are acquired, particularly in an EFL environment, as recommended by Cardoso, Smith, and Garcia Fuentes (2015).

The objective of this study is to evaluate the voice quality of a standard TTS system in comparison with that of a human. It also examines the pedagogical potential of using TTS-based input in an EFL setting in terms of its speech quality, ability to be understood by L2 users, and potential to focus on specific language features according to the following criteria:

- text comprehension (an intelligibility test to assess users' ability to understand a text and answer comprehension questions);
- intelligibility (the extent to which a message is actually understood, measured by a dictation activity; Derwing & Munro, 2005);
- users' ratings of holistic pronunciation features (comprehensibility, naturalness and accuracy; Derwing & Munro, 2005); and
- users' ability to identify a linguistic feature (i.e. the aural identification of English regular past tense endings).

This study is guided by the following research question: how does the quality of speech produced by a TTS system compare to a human voice?

2. Method

Twenty-nine adult Brazilian EFL learners (native speakers of Brazilian Portuguese) were recruited in Recife, Brazil (age range: 18-33; M=23.6, SD=4.9). Their proficiency in English was intermediate, determined by a triangulation of methods (placement at their language school, self-ratings, and the researcher's assessment during the experiment).

Data were collected in one-shot individual sessions in which each participant completed a set of tasks designed to assess each criterion pertinent to evaluating the quality of TTS and human speech. For intelligibility, participants completed a 'dictation task' in which they were asked to transcribe sentences. In addition, they listened to two short stories and answered six multiple-choice questions covering

each story's main points. To evaluate pronunciation holistically, participants rated the quality of speech based on comprehensibility, naturalness, and pronunciation accuracy using a six point Likert scale. Participants rated not only the two short stories, but also 12 decontextualised short sentences (e.g. 'The boy watched the clock ticking on the wall'). The rationale for the inclusion of these decontextualised sentences was that they could yield different results due to the low cognitive load required for their processing, as the participants need to concentrate solely on speech quality, not understanding. Finally, for the ability to focus on grammatical forms, participants performed an aural identification task for 16 sentences, in which they judged whether the target feature (past tense marker -ed) appeared in the input or not. Participants had to determine whether the action took place in the past (e.g. 'I called my mother') or not (e.g. 'I visit my cousin Sam') and check their response on the answer sheet.

For all tasks, participants listened to speech samples alternately produced by TTS and a human. The TTS voice, by NeoSpeech, was based on a female North American speaker, and the human was a North American female native-speaker with similar speech patterns. The material presented to participants was organised in two randomised sequences (A, B) in a way that both sequences contained the same target sentences or texts, but were produced by different voice sources. Participants who received Sequence A heard the same sentences as participants in Sequence B; however, all the sentences produced by the TTS in Sequence A were recorded by human voice in Sequence B, and vice-versa. At the end of the session, participants were interviewed about their insights on the quality of the TTS-generated voices. This paper reports the findings from the analysis of the quantitative data gathered from the four tasks.

3. Results

Participants' ratings of speech quality (comprehensibility, naturalness and accuracy) in short stories and sentences, story comprehension results, percentage of correct words transcribed in the dictation task (intelligibility), and participants' accuracy in identifying present/regular past verbs (to measure TTS's ability to provide noticeable input) were tallied and the means of matched voice pairings from both randomised sequences (A, B) were compared. Parametric statistics were used for data sets that met normality assumptions (namely data from the short story comprehension tasks and ratings). For every other set, non-parametric tests were conducted. Paired sample t-tests and Wilcoxon Signed-Rank tests were used, respectively, with an alpha level of .05 to determine statistical significance. An

adjusted alpha of .004 was calculated using a false detection rate post-hoc method. Table 1, Table 2, and Table 3 below show the descriptive statistics and results according to each task.

Table 1. Descriptive statistics and parametric results: story rating, story comprehension (intelligibility)

	TTS		**Human**		**t**	**p**
Story ratings	Mean	SD	Mean	SD		
Comprehensibility	4.42	.02	4.92	.30	-2.59	.235
Naturalness	3.12	.74	4.58	.41	-6.35	.099
Accuracy	5.04	.15	5.31	.13	-27.00	.024
Comprehension test (intelligibility)	4.57	.81	4.74	.75	-4.25	.147

Table 2. Descriptive statistics and nonparametric results: sentence rating, dictation (intelligibility)

	TTS	**Human**	**Z**	**p**
Sentence ratings	Median	Median		
Comprehensibility	5.06	5.10	-.628	.530
Naturalness	3.45	5.13	-3.06	.002*
Accuracy	4.93	5.10	-2.85	.004*
Dictation task (Intelligibility)	59.65	55.05	-.153	.878

*p< .004

Table 3. Descriptive statistics and nonparametric test results for feature identification test

	Median	**Z**	**p**
TTS	.67	-1.67	.094
Human	.83		

The statistical analyses showed that foreign language learners rated or performed similarly regardless of the voice (TTS or human), except for naturalness and accuracy at sentential levels. The findings correspond to previous studies (e.g. Cardoso et al., 2015) and to those obtained in Kang, Kashiwagi, Treviranus, and Kaburagi (2008), who found that non-native English learners do not recognise a significant difference between synthetic and human voices. Contrary to previous studies, such as Bailly (2003), this study found that artificial and human speech were equally intelligible and comprehensible. Finally, echoing the results of Cardoso et al. (2015), both TTS and human samples helped participants notice

past tense forms, confirming our hypothesis that TTS can provide learners with alternative ways to access or identify target linguistic forms.

4. Conclusions

As recommended by Cardoso et al. (2015), evaluations of TTS systems should be conducted in EFL environments where language exposure is limited to determine their effectiveness as pedagogical tools in providing students with additional opportunities for practice. The speech synthesis evaluated in this study has generally performed equally to a human voice, demonstrating a high level of intelligibility and the ability to provide learners with opportunities to notice aural linguistic forms such as the regular past -ed. This finding indicates that a change in learning environment (from second to foreign) can positively affect learners' perceptions and attitudes towards TTS-produced input, and suggests that EFL learners may be less sensitive to distinctions between natural and artificial voices than ESL students. Future research should reinforce these results by evaluating TTS in other EFL settings to verify if students in these contexts could also benefit from its adoption.

Our findings suggest that TTS systems are ready for L2 pedagogy, as they can enhance learners' access to the target language anytime and anywhere, promote autonomous learning (e.g. where students select their own materials), and facilitate teacher-supervised instruction (e.g. where teachers develop personalised materials for their students, based on their needs).

References

Bailly, G. (2003). Close shadowing natural versus synthetic speech. *International Journal of Speech Technology, 6*(1), 11-19. https://doi.org/10.1023/A:1021091720511

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. https://doi.org/10.14705/rpnet.2015.000318

Derwing, T. M., & Munro, M. J. (2005). Second language accent and pronunciation teaching: a research-based approach. *TESOL Quarterly, 39*(3), 379-397. https://doi.org/10.2307/3588486

Kang, M., Kashiwagi, H., Treviranus, J., & Kaburagi, M. (2008). Synthetic speech in foreign language learning: an evaluation by learners. *International Journal of Speech Technology, 11*(2), 97-106. https://doi.org/10.1007/s10772-009-9039-3

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

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. (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), 71-93.

Research-publishing.net

Published by Research-publishing.net, not-for-profit association
Contact: info@research-publishing.net

CALL in a climate of change: adapting to turbulent global conditions – short papers from EUROCALL 2017
Edited by Kate Borthwick, Linda Bradley, and Sylvie Thouësny

Typeset by Research-publishing.net

Cover design based on © Josef Brett's, Multimedia Developer, Digital Learning, http://www.eurocall2017.uk/, reproduced with kind permissions from the copyright holder.

Cover layout by © Raphaël Savina (raphael@savina.net)
Photo "frog" on cover by © Raphaël Savina (raphael@savina.net)

Fonts used are licensed under a SIL Open Font License

ISBN13: 978-2-490057-04-7 (Ebook, PDF, colour)
ISBN13: 978-2-490057-05-4 (Ebook, EPUB, colour)
ISBN13: 978-2-490057-03-0 (Paperback - Print on demand, black and white)
Print on demand technology is a high-quality, innovative and ecological printing method; with which the book is never 'out of stock' or 'out of print'.

British Library Cataloguing-in-Publication Data.
A cataloguing record for this book is available from the British Library.

Legal deposit: Bibliothèque Nationale de France - Dépôt légal: décembre 2017.