Evaluating text-to-speech synthesizers

Walcir Cardoso[1], George Smith[2], and Cesar Garcia Fuentes[3]

Abstract. Text-To-Speech (TTS) synthesizers have piqued the interest of researchers for their potential to enhance the L2 acquisition of writing (Kirstein, 2006), vocabulary and reading (Proctor, Dalton, & Grisham, 2007) and pronunciation (Cardoso, Collins, & White, 2012; Soler-Urzua, 2011). Despite their proven effectiveness, there is a need for up-to-date formal evaluations of TTS systems. The present study was an attempt to evaluate the language learning potential of an up-to-date TTS system at two levels: (1) speech quality (comprehensibility, naturalness, accuracy, and intelligibility) and (2) focus on a linguistic form (via a feature identification task). For Task 1, participants listened to and rated human- and TTS-produced stories and sentences on a 6-point scale (1); for Task 2, they listened to 16 human- and TTS-produced sentences to identify the presence of a target feature (English regular past -ed). Results of paired samples t-tests indicated that for speech quality, the human samples earned higher ratings than the TTS samples. For the second task (past -ed perception), the TTS and human-produced samples were equivalent. The discussion of the findings will highlight how TTS can be used to complement and enhance the teaching of L2 pronunciation and other linguistic skills both inside and outside the classroom.

Keywords: computer-assisted language learning, CALL, text-to-speech, technology and language learning.

1. Concordia University, Canada; walcir@education.concordia.ca
2. University of Hawaii at Manoa, United States; gfsmith@hawaii.edu
3. Concordia University, Canada; cesgarfu@hotmail.com

How to cite this article: 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). Dublin: Research-publishing.net. http://dx.doi.org/10.14705/rpnet.2015.000318

1. Introduction

The provision of target language input of sufficient quality and quantity is an important issue in the field of second language acquisition. Three challenges which exist with this provision are: (1) the need for vast amounts of comprehensible input to develop language competence (Council of Europe, 2001; Krashen, 1985); (2) the need for learner-centered and personalized input (Chapelle, 2001); and (3) the need for exposure to a variety of speech models for robust phonological development (Barcroft & Sommers, 2005).

Traditional face-to-face classroom settings may not be able to meet these criteria due to the inherent restrictions of this teaching context (e.g. teacher-centered, one variety of English used, lack of sustained input practice), especially in foreign-language settings (Cardoso et al., 2012). One remedy to this problem lies in the use of TTS, which can offset some of the limitations of traditional classrooms given that they are highly flexible, learner-centered, and easily accessible. Several studies have attested to the benefits of using TTS for learning writing (Kirstein, 2006), vocabulary and reading (Proctor et al., 2007) and pronunciation (Cardoso et al., 2012; Soler-Urzua, 2011), both in and outside the classroom.

Despite these theoretical and empirical benefits, however, there exist very few formal evaluations of TTS systems, specifically of their potential to promote the ideal conditions under which Second Language Learning (SLA) is thought to occur – a critical stage in the evaluation of Computer-Assisted Language Learning (CALL) applications (Chapelle, 2001; Handley & Hamel, 2005). Those evaluations which do exist have used a wide variety of rating methods, produced mixed results (some demonstrating the adequacy of TTS – Kang, Kashiwagi, Treviranus, & Kaburagi, 2009, some demonstrating inadequacy in some respects – Handley, 2009, Nusbaum, Francis, & Henley, 1995) and date back more than 5 years. The present study was thus an attempt to provide an up-to-date evaluation of a state of the art TTS system concerning its potential to promote ideal SLA processes. The following two evaluation criteria were chosen: (1) the speech quality of the TTS system (input); and (2) the potential for learners to focus on linguistic form in these two types of input. Two research questions were formulated as follows:

- What is the quality of speech produced by TTS systems in comparison with that by humans?

- Can TTS systems provide learners with the opportunity to focus on form?

2. Method

2.1. Participants and design

Fifty-four university-level participants with a variety of L1 backgrounds were recruited at an English-language university in Canada. Two tasks were designed to elicit participants' perceptions of the TTS system: rating speech quality and feature identification. Both tasks had learners listen to speech samples produced by TTS and a human, with the goal being human-TTS equivalency; accordingly, a paired samples design was adopted for the analysis.

2.2. Stimuli and materials

A female speaker of North American English (Julie) in the program NaturalReader 13 (2013) was used as the TTS system, and compared with a native-speaker of the same dialect with similar speech properties. For speech quality, participants listened to two stories and twelve sentences and rated them according to four judgment criteria: comprehensibility, naturalness, pronunciation accuracy, and intelligibility on a 6-point scale. Potential for focus on form was measured by having learners perform an aural feature identification task wherein they judged whether certain sentences contained a target grammar feature (English regular past -ed). Both the stories and sentences were adapted from materials produced by the ALERT research project (Collins et al., 2011). The tasks were performed via Microsoft PowerPoint in a quiet lab at the university, by a trained research assistant.

2.3. Analysis

Data came from the participants' judgments of the stories and sentences that they heard and their accuracy on perceiving past -ed in decontextualized sentences such as "I hated the movie" and "I hate the movie"). The ratings of each participant were tallied, and means were calculated for each story and sentence. Accuracy scores were reported as raw scores, with a maximum of 8 points per speech source (i.e. human or TTS). Main analysis was carried out by means of paired samples t-tests, with an alpha level of .05 used for the determination of statistical significance.

3. Results

Results for the rating task were as follows. For the stories, paired samples t-tests revealed a significant difference in the rating scores on all categories (comprehensibility, $t(54)=-4.77, p<.001$; naturalness, $t(54)=-9.35, p<.001$; accuracy,

t(54)=-7.32, *p*<.001; and intelligibility, *t*(54)=-6.40, *p*<.001). Similarly, paired samples t-tests for the sentences also revealed significant differences between the human- and TTS-produced samples for all measures (comprehensibility, *t*(54)=-6.13, *p*<.001; naturalness, *t*(54)=-7.63, *p*<.001; accuracy, *t*(54)=-7.34, *p*<.001; and intelligibility, *t*(54)=-6.11, *p*<.001). For the past -ed identification task, paired samples t-tests revealed no significant differences between the TTS- and human-produced speech samples (*t*(54)=-1.93, *p*=.059). Table 1, Table 2, and Table 3 below show the descriptive statistics according to each task.

Table 1. Descriptive statistics for story rating

	Comprehensibility		Naturalness		Accuracy		Intelligibility	
	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
Human	5.66	0.67	5.61	0.68	5.82	0.47	5.54	0.79
TTS	5.14	0.94	3.64	1.63	4.77	1.06	4.50	1.18

Table 2. Descriptive statistics for sentence rating

	Comprehensibility		Naturalness		Accuracy		Intelligibility	
	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
Human	5.90	0.29	5.65	0.35	5.74	0.29	5.80	0.38
TTS	5.36	0.69	4.24	1.50	5.35	0.83	4.94	1.01

Table 3. Descriptive statistics for feature identification task

	Mean	*SD*
Human	6.03	1.38
TTS	5.59	1.12

4. Discussion and conclusions

The present study sought to evaluate the speech quality and potential to focus on linguistic form provided by a state-of-the-art TTS system. First, the results revealed that the samples produced by the TTS system were rated significantly lower than the human-produced samples for all four categories of speech quality (comprehensibility, naturalness, pronunciation accuracy, intelligibility), at both the story and sentence levels. This echoes previous findings that have shown less favorable ratings for TTS-produced speech compared to human speech (e.g. Handley, 2009; Handley & Hamel, 2005; Nusbaum et al., 1995). However, it is

important to observe that the mean rating scores assigned to the TTS system for 3 out of the 4 categories (naturalness excluded) were relatively high (4.5-5.3 out of 6). Thus, the speech quality of this particular TTS system can be considered as having achieved the "top rating(s)" needed for advancement to the next stage of evaluation (i.e. the success of activities using TTS) and use in language learning in general (Handley, 2009). The results of the past -ed perception task offer similarly promising results. Statistical equivalency was found for participants' ability to detect the presence of the target feature (past -ed) with high accuracy (~5.5 or 6 out of 8). This indicates that regardless of the source of delivery (human or TTS), participants were equally able to perceive the target form in running speech.

Implications of these results are that modern TTS systems seem to be ready for advancement to further stages of evaluation, but more importantly, for use in language learning activities, particularly as a supplemental source of input which can cater to learners' individual needs and interests. Future research should not only undertake evaluations of TTS' success as a learning tool in classrooms (particularly in English as a foreign language classrooms, where language exposure is limited), but also continue evaluations for a variety of other factors, such as the level of cognitive processing involved in listening to computer-generated speech.

5. Acknowledgements

We would like to thank the participants, Fatma Bouhlal, and Suzanne Ceretta (the human voice) for their invaluable contributions to this project. We would also like to acknowledge the financial support from the Social Sciences and Humanities Research Council of Canada.

References

Barcroft, J., & Sommers, M. S. (2005). Effects of acoustic variability on second language vocabulary learning. *Studies in Second Language Acquisition, 27*, 387-414. doi:10.1017/s0272263105050175

Cardoso, W., Collins, L., & White, J. (2012). Phonological input enhancement via text-to-speech synthesizers: the L2 acquisition of English simple past allomorphy. *Paper presented at the American Association of Applied Linguistics conference, Boston, U.S.A*.

Chapelle, C. A. (2001). Innovative language learning: achieving the vision. *ReCALL, 13*(1), 3-14. doi:10.1017/S0958344001000210

Collins, L., Horst, M., Trofimovich, P., White, J., & Cardoso, W. (2011). Explaining and enhancing efficiency in classroom second language learning. *Research Grant from the Fonds québécois de la recherché sur la société et la Culture (FQRSC), Soutiens aux équipes de recherche*.

Council of Europe (2001). Common European Framework of Reference for Languages: Learning, teaching, assessment. Cambridge: Cambridge University Press.

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

Handley, Z., & Hamel, M. J. (2005). Establishing a methodology for benchmarking speech synthesis for computer-assisted language learning (CALL). *Language Learning & Technology, 9*(3), 99-120.

Kang, M., Kashigawi, H., Treviranus, J., & Kaburagi, M. (2009). Synthetic speech in foreign language learning: an evaluation by learners. *International Journal of Speech Technology, 11*(2), 97-106. doi: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, U.S.A.

Krashen, S. (1985). The input hypothesis: issues and implications. New York: Longman.

Nusbaum, H. C., Francis, A. L., & Henly, A. S. (1995). Measuring the naturalness of synthetic speech. *International Journal of Speech Technology, 1*(1), 7-19. doi:10.1007/BF02277176

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.

Soler-Urzua, F. (2011). *The acquisition of English /ɪ/ by Spanish speakers via text-to-speech synthesizers: a quasi-experimental study*. Master's Thesis. Concordia University, Montreal, Canada.

Published by Research-publishing.net, not-for-profit association
Dublin, Ireland; info@research-publishing.net

Critical CALL – Proceedings of the 2015 EUROCALL Conference, Padova, Italy
Edited by Francesca Helm, Linda Bradley, Marta Guarda, and Sylvie Thouësny

Typeset by Research-publishing.net
Fonts used are licensed under a SIL Open Font License

ISBN13: 978-1-908416-28-5 (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'.

ISBN13: 978-1-908416-29-2 (Ebook, PDF, colour)
ISBN13: 978-1-908416-30-8 (Ebook, EPUB, colour)

Legal deposit, Ireland: The National Library of Ireland, The Library of Trinity College, The Library of the University of Limerick, The Library of Dublin City University, The Library of NUI Cork, The Library of NUI Maynooth, The Library of University College Dublin, The Library of NUI Galway.

Legal deposit, United Kingdom: The British Library.
British Library Cataloguing-in-Publication Data.
A cataloguing record for this book is available from the British Library.

Legal deposit, France: Bibliothèque Nationale de France - Dépôt légal: décembre 2015.