Who's got talent? Comparing TTS systems for comprehensibility, naturalness, and intelligibility

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

Abstract. The current study compared five free Text-To-Speech (TTS) systems, selected based on characteristics such as availability and capabilities. Tasks were completed by 37 English learners to evaluate these systems in terms of their comprehensibility, naturalness, and intelligibility. Our findings indicate that IBM Watson and Google Translate are the best TTS systems, according to the evaluation criteria employed.

Keywords: text-to-speech synthesis, L2 pronunciation.

1. Introduction

Second language (L2) researchers have explored the pedagogical capabilities of TTS synthesizers for their potential to enhance the acquisition of writing (Kirstein, 2006), vocabulary, and reading (Proctor, Dalton, & Grisham, 2007), and pronunciation (e.g. Liakin, Cardoso, & Liakina, 2017). In addition, recent evaluations of TTS quality have attested that students perceive little difference between synthetic and human speech (Bione Alves, 2017; Cardoso, Smith, & Garcia Fuentes, 2015), suggesting that TTS technology is ready for pedagogical use not only because of its voice quality, but also because users may have become familiar with synthesized speech.

The availability of free web-based TTS applications is also promising (Karakaş, 2017), as L2 students can access these tools from any device. However, faced with

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

How to cite this article: Grimshaw, J., Bione, T., & Cardoso, W. (2018). Who's got talent? Comparing TTS systems for comprehensibility, naturalness, and intelligibility. In P. Taalas, J. Jalkanen, L. Bradley & S. Thouësny (Eds), *Future-proof CALL: language learning as exploration and encounters – short papers from EUROCALL 2018* (pp. 83-88). Research-publishing.net. https://doi.org/10.14705/rpnet.2018.26.817

a plethora of options, users may find it difficult to choose the most appropriate TTS system to use, particularly in terms of voice quality. As these technologies evolve, there is a need for regular evaluations to determine which systems will best suit L2 users' needs.

The current study compared a set of five TTS systems in terms of their comprehensibility, naturalness, and intelligibility, based on tasks completed by a group of English as a Second and Foreign Language (ESL/EFL) learners. It was guided by the following research question: which of the selected five TTS systems constitute the most pedagogically appropriate software in terms of their ability to produce speech that is comprehensible, natural, and intelligible?

2. Method

While seven freely available TTS systems were originally considered for analysis, we decided to select a more manageable number of TTS systems for evaluation based on criteria that included: availability (web, iOS, Android), popularity (ratings on Google Play or the App Store), and other pedagogically-relevant capabilities (e.g. ability to control voice speed and pitch; see Barcroft & Sommers, 2005 for the rationale). Consequently, the number of TTS systems for evaluation was reduced to five: IBM Watson, Google Translate, LumenVox, NeoSpeech, and NaturalReader.

2.1. Participants

Participants were 37 native speakers of Brazilian Portuguese (17 males, 20 females; all adults), living in Brazil (EFL; n=12) or abroad (ESL; n=25). The study targeted high-intermediate to advanced learners (self-prescribed) recruited over social media. For the purposes of the current analysis, results from both ESL and EFL groups were combined.

2.2. Instruments

For each of the five TTS systems, one short story clip (20-30 seconds) and four short sentences were recorded using Audacity (or downloaded from the TTS application, if the option was available). Only default female voices were used in the recordings, as not all TTS systems offered male voices; all other default settings were also retained (speed, pitch, etc.) to replicate a user attempting to use the system without guidance. All recordings were placed into a quiz on a Moodle-based testing environment.

2.3. Procedures

After participants gave their consent, they first practiced a set of ratings before completing a Moodle-based test, all of which was conducted remotely (online); the process lasted approximately 15 minutes. The participants were instructed to listen to each recording only once.

To evaluate the selected TTS systems in terms of comprehensibility and naturalness, participants listened to and rated recordings of short story clips from five different systems. They were also asked to transcribe four sentences produced by each TTS system (for a total of 20) as a measure of intelligibility.

2.4. Data collection and analysis

Participants rated TTS voices based on a six-point Likert scale for comprehensibility (1=very difficult to understand, 6=very easy to understand) and naturalness (1=very unnatural, 6=very natural). For these items, descriptive statistics (means, standard deviations) were reported. For intelligibility, the data were measured according to transcription accuracy percentage, where participants could transcribe between 0% to 100% of each sentence correctly. It was assumed that more intelligible sentences would result in more accurately transcribed words.

3. Results

3.1. Comprehensibility

Participant comprehensibility ratings suggest that IBM's Watson TTS system is the most favorable (M=5.81, SD=0.46), followed by Google Translate (M=5.35, SD=0.82); see Table 1 for all ratings.

Table 1. Comprehensibility ratings

TTS system	Mean	Standard deviation
IBM Watson	5.81	0.46
Google Translate	5.35	0.82
LumenVox	4.81	1.02
NeoSpeech	4.41	1.01
NaturalReader	3.78	1.46

3.2. Naturalness

Participant ratings for naturalness follow a similar trend as comprehensibility, with Watson being ranked as most natural (*M*=4.87, *SD*=1.03), followed by Google Translate (*M*=3.73, *SD*=1.46); see Table 2.

Table 2. Naturalness ratings

TTS system	Mean	Standard deviation
IBM Watson	4.87	1.03
Google Translate	3.73	1.46
NeoSpeech	3.14	1.25
LumenVox	3.08	1.19
NaturalReader	2.03	0.96

3.3. Intelligibility

Accuracy percentages for the intelligibility task (Table 3) indicate that Watson's voice once again scored the highest (*M*=90%, *SD*=5%), followed by LumenVox (*M*=88%, *SD*=10%).

Table 3. Intelligibility accuracy scores

TTS system	Mean	Standard deviation
IBM Watson	90%	5%
LumenVox	88%	10%
Google Translate	85%	7%
NaturalReader	85%	14%
NeoSpeech	79%	16%

4. Discussion and concluding remarks

According to participant ratings, IBM Watson appears to be the most comprehensible and natural, followed by Google Translate. In terms of intelligibility, accuracy scores suggest that, once again, IBM Watson comes out on top, followed by LumenVox. One reason why IBM Watson outranks all others in these measures may be because it is a demo version of a new and highly advanced system which highlights the capabilities of state-of-the-art TTS (e.g. users can modify the voice's expression or pitch to make the voice sound apologetic or anxious; see demo: https://text-to-speech-demo.ng.bluemix.net). Google Translate may have also received high ratings because its synthesized voice is commonly used in many popular apps and

websites. As a result, it is possible that participants may have already been familiar with the voice adopted in our study. Bione Alves (2017), for instance, noted that users' rating for comprehensibility and naturalness might increase as participants become more acquainted with synthetic voices.

There are several reasons that may explain the lower ratings for the other systems. As NaturalReader's default settings had the TTS voice play at a higher than normal speed, this may have influenced rater comprehensibility and naturalness. In the free version of NeoSpeech, soft music plays in the background as the voice speaks; this may have therefore interfered with user comprehensibility, while the creators of LumenVox may have focused on the quantity of voices they offer rather than quality. As we also only targeted one voice per system, ratings for the five systems may have varied if different voices had been used. Additionally, in reality, many TTS applications (including IBM Watson, Google Translate, NaturalReader, NeoSpeech) have user-controlled features that allow them to modify the speed of speech and/or repeat the speech as many times as necessary, a feature that may place some of these TTS systems at the same level, considering the three criteria adopted to evaluate them.

To conclude, the aim of this study was to evaluate and compare five TTS systems in terms of their comprehensibility, naturalness, and intelligibility, as assessed by a group of ESL/EFL learners. The results obtained in our analysis of participants' ratings and transcriptions suggest that, among the TTS systems considered, IBM Watson and Google Translate constitute, at present, the more pedagogically appropriate choices for L2 learners willing to enhance (in both quantity and quality) their access to the target language.

Assuming that the pedagogical use of TTS has the potential to extend the reach of the classroom (Bione Alves, 2017; Cardoso et al., 2015) and that it is beneficial for learning (Liakin et al., 2017), teachers can use one of these readily available systems to develop activities and tasks to provide additional listening and pronunciation practice. Although TTS systems are undergoing constant change, we hope that the criteria outlined here will help the language teacher to critically select the most pedagogically appropriate tool for their purposes.

5. Acknowledgements

We would like to thank our participants and Paul John for his input. This project was partially funded by the *Social Sciences and Humanities Research Counsel of Canada*.

References

Barcroft, J., & Sommers, M. S. (2005). Effects of acoustic variability on second language vocabulary learning. *Studies in Second Language Acquisition, 27*(3), 387-414. https://doi.org/10.1017/S0272263105050175

Bione Alves, T. (2017). *Synthetic voices in the foreign language context*. Master's Thesis. Concordia University, Montreal, CA.

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

Karakaş, A. (2017). English voices in 'text-to-speech tools': representation of English users and their varieties from a World Englishes perspective. *Advances in Language and Literary Studies, 8*(5), 108-119. https://doi.org/10.7575/aiac.alls.v.8n.5p.108

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

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), 71-79.

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

Future-proof CALL: language learning as exploration and encounters – short papers from EUROCALL 2018
Edited by Peppi Taalas, Juha Jalkanen, Linda Bradley, and Sylvie Thouësny

Publication date: 2018/12/08

Typeset by Research-publishing.net
Cover theme by © 2018 Antti Myöhänen (antti.myohanen@gmail.com)
Cover layout by © 2018 Raphaël Savina (raphael@savina.net)
Drawings by © 2018 Linda Saukko-Rauta (linda@redanredan.fi)

ISBN13: 978-2-490057-22-1 (Ebook, PDF, colour)
ISBN13: 978-2-490057-23-8 (Ebook, EPUB, colour)
ISBN13: 978-2-490057-21-4 (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, UK: British Library.
Legal deposit, France: Bibliothèque Nationale de France - Dépôt légal: Décembre 2018.