Assessing the quality of TTS audio in the LARA learning-by-reading platform

Elham Akhlaghi[1], Anna Bączkowska[2], Harald Berthelsen[3], Branislav Bédi[4], Cathy Chua[5], Catia Cucchiarini[6], Hanieh Habibi[7], Ivana Horváthová[8], Pernille Hvalsøe[9], Roy Lotz[10], Christèle Maizonniaux[11], Neasa Ní Chiaráin[12], Manny Rayner[13], Nikos Tsourakis[14], and Chunlin Yao[15]

Abstract. A popular idea in Computer Assisted Language Learning (CALL) is to use multimodal annotated texts, with annotations typically including embedded audio and translations, to support L2 learning through reading. An important question is how to create the audio, which can be done either through human recording or by a Text-To-Speech (TTS) synthesis engine. We may reasonably expect TTS to be quicker and easier, but humans to be of higher quality. Here, we report a study using the open-source LARA platform and ten languages. Samples of LARA audio totaling about three and a half minutes were provided for each language in both human and TTS form; subjects used a web form to compare different versions of the same item and rate the voices as a whole. Although human voice was more often preferred, TTS achieved higher ratings in some languages and was close in others.

Keywords: reading, multimodality, TTS, evaluation.

1. Ferdowsi University of Mashhad, Mashhad, Iran; elham.akhlaghi@mail.um.ac.ir
2. University of Gdansk, Gdansk, Poland; anna.baczkowska@ug.edu.pl
3. Trinity College, Dublin, Ireland; berthelh@tcd.ie
4. The Árni Magnússon Institute for Icelandic Studies, Reykjavík, Iceland; branislav.bedi@arnastofnun.is
5. Independent scholar, Adelaide, Australia; cathyc@pioneerbooks.com.au
6. Radboud University, Nijmegen, The Netherlands; c.cucchiarini@let.ru.nl
7. University of Geneva, Geneva, Switzerland; hanieh.habibi@unige.ch
8. Univerzita Konstantina Filozofa, Nitra, Slovakia; ihorvathova@ukf.sk
9. University of Copenhagen, Copenhagen, Denmark; phv@hum.ku.dk
10. Independent scholar, Madrid, Spain; roylotz09@hotmail.com
11. Flinders University, Adelaide, Australia; christele.maizonniaux@flinders.edu.au
12. Trinity College, Dublin Ireland; neasa.nichiarain@tcd.ie
13. University of Geneva, Geneva, Switzerland; emmanuel.rayner@unige.ch
14. University of Geneva, Geneva, Switzerland; nikolaos.tsourakis@unige.ch
15. Tianjin Chengjian University, Tianjin, China; yao_chunlin@126.com

How to cite this article: Akhlaghi, E., Bączkowska, A., Berthelsen, H., Bédi, B., Chua, C., Cucchiarini, C., Habibi, H., Horváthová, I., Hvalsøe, P., Lotz, R., Maizonniaux, C., Ní Chiaráin, N., Rayner, M., Tsourakis, N., & Yao, C. (2021). Assessing the quality of TTS audio in the LARA learning-by-reading platform. In N. Zoghlami, C. Brudermann, C. Sarré, M. Grosbois, L. Bradley, & S. Thouësny (Eds), *CALL and professionalisation – short papers from EUROCALL 2021* (pp. 1-5). Research-publishing.net. https://doi.org/10.14705/xxx-to-be-confirmed

1. Introduction

An increasingly popular idea over the last decade is to help L2 learners improve their reading skills in non-native languages by creating annotated multimedia texts that contain integrated help, most commonly word translations and/or audio. High profile examples include LingQ[16] and Learning With Texts[17].

In this study, our focus is the audio, created either by recording human voice or through a TTS engine. Using TTS is faster, but despite on-going improvements in TTS technology, human-recorded audio is still of higher voice quality. It is less clear how large the difference is, or how important it is in practice when TTS is used in L2 teaching. Our study addresses these questions.

2. Method

The experiments were performed using LARA[18] (Akhlaghi et al., 2020), a learning-by-reading platform under development by an international open-source consortium since 2018. So far, most LARA texts have used human-recorded audio, though the Irish LARA group has consistently used TTS. In our study, we selected existing LARA texts in Danish, English, Farsi, French, Icelandic, Irish, Italian, Mandarin, Spanish, and Swedish, creating a version using the other method so that it was available in both human and TTS voice. For each language, a single human voice was used and TTS audio was created using the best TTS engine available to us for that language: ABAIR[19] for Irish, Google TTS[20] for Mandarin, Nuance Vocalizer[21] for Farsi, and ReadSpeaker[22] for the other languages.

We randomly selected contiguous passages from the texts so that the total audio for each language was about three and a half minutes; for some languages, we also included individual words. The material was presented on an openly available anonymous web form consisting of three portions: demographic data; item-by-item comparison of the audio; and overall impressions of the two voices. In the item-by-item comparison, subjects chose between 'both acceptable and roughly equal', 'both acceptable but one clearly better', 'one acceptable, one not acceptable', and 'neither acceptable'. In the overall impressions part,

16. https://www.lingq.com/
17. https://sourceforge.net/projects/lwt/
18. https://www.unige.ch/callector/lara/
19. http://www.abair.ie/
20. https://cloud.google.com/text-to-speech
21. https://www.nuance.com/en-au/omni-channel-customer-engagement/voice-and-ivr/text-to-speech/vocalizer.html
22. https://www.readspeaker.com/

subjects gave Likert scale scores for quality of individual words, quality of whole sentences, speed, naturalness, pleasantness, suitability for teaching, suitability for imitating, and a freeform response. Full details are posted in the supplementary materials.

3. Results

Responses were logged for 130 subjects and collated using a script. There were large differences between languages, between responses for sentences and words, and between native and non-native judgments. Table 1, Table 2, Table 3, and Table 4 show results for the portions of the data we considered most informative. Full details are posted in the supplementary materials.

Table 1. Overall impressions of voices, five-point Likert scale; ratings from **native/near-native speakers** only. In each cell, human rating above and TTS rating below. Yellow=TTS equal or better than human, orange=TTS within 0.5 of human

Language	DA	EN	FA	FR	IS	IE	IT	ZH	SP	SW
(#raters)	(7)	(13)	(27)	(3)	(6)	(6)	(7)	(2)	(2)	(3)
Words	3.86 3.71	3.77 4.08	4.74 3.19	4.0 5.0	4.17 3.17	4.5 4.5	3.43 4.43	4.0 5.0	4.5 2.5	4.67 4.33
Sentences	4.29 2.57	3.77 3.77	4.78 2.74	4.33 4.67	3.83 3.17	4.67 4.33	4.14 3.29	4.5 3.5	4.5 2.0	4.33 4.33
Speed	4.14 2.57	3.62 3.92	4.78 3.3	5.0 4.67	3.83 3.67	4.67 3.67	4.14 3.71	4.5 3.5	4.5 3.5	4.67 4.33
Natural	4.29 1.86	3.85 3.46	4.85 2.22	5.0 4.0	4.5 3.0	4.83 3.5	4.86 1.86	5.0 3.0	3.5 1.5	4.67 4.33
Pleasant	4.14 2.43	3.46 3.62	4.63 2.41	4.33 4.33	4.33 3.33	4.83 3.33	4.14 2.43	5.0 3.5	3.5 2.5	4.33 4.33
Teaching	4.43 2.43	3.54 3.62	4.78 2.33	4.0 4.33	4.33 3.5	4.33 3.5	3.43 3.14	5.0 3.5	4.5 2.0	4.0 4.33
Imitating	4.43 2.14	3.0 3.54	4.7 2.19	4.33 4.67	3.83 2.83	4.67 3.17	3.57 2.29	5.0 3.5	4.5 1.5	4.0 4.0

Table 2. Item-by-item comparison averages; percentage ratings from **native/near-native speakers** only, **sentences** only. Yellow=TTS equal or better than human, orange=TTS within 10% of human

Language	DA	EN	FA	FR	IS	IE	IT	ZH	SP	SW
(#raters)	(7)	(13)	(27)	(3)	(6)	(6)	(7)	(2)	(2)	(3)
(#items)	(14)	(8)	(20)	(22)	(15)	(39)	(23)	(17)	(16)	(14)

Human acceptable	98.0	100.0	98.3	84.8	94.4	97.9	88.8	100.0	100.0	97.6
TTS acceptable	41.8	88.5	66.5	97.0	81.1	99.6	85.1	58.8	40.6	97.6
Human better	81.6	43.3	70.9	6.1	53.3	47.0	31.7	47.1	100.0	40.5
TTS better	3.1	20.2	6.1	27.3	14.4	7.7	25.5	2.9	0.0	7.1
(same)	(15.3)	(36.5)	(23.0)	(66.7)	(32.2)	(45.3)	(42.9)	(50.0)	(0.0)	(52.4)

Table 3. Overall impressions of voices; **teachers/trainee teachers** only (conventions as in Table 1)

Language	**DA**	**EN**	**FA**	**FR**	**IS**	**IE**	**IT**	**ZH**	**SP**	**SW**
(#raters)	(7)	(16)	(13)	(4)	(9)	(25)	(2)	(2)	(1)	(1)
Words	3.86 3.71	3.62 4.25	4.69 3.15	4.5 5.0	4.33 3.22	4.2 3.8	3.5 4.0	4.0 5.0	5.0 2.0	5.0 3.0
Sentences	4.29 2.57	3.5 4.0	4.85 2.92	4.5 4.5	4.22 3.22	4.24 3.76	3.0 3.5	4.5 3.5	5.0 2.0	5.0 3.0
Speed	4.14 2.57	3.25 4.0	4.85 3.15	4.25 4.25	4.22 3.67	4.2 3.2	3.5 3.5	4.5 3.5	5.0 4.0	4.0 4.0
Natural	4.29 1.86	3.25 3.94	5.0 2.46	5.0 3.5	4.67 2.78	4.2 4.0	4.5 1.5	5.0 3.0	5.0 1.0	5.0 3.0
Pleasant	4.14 2.43	2.88 3.88	4.77 2.85	5.0 3.0	4.44 3.22	4.28 3.16	3.5 2.5	5.0 3.5	5.0 2.0	4.0 4.0
Teaching	4.43 2.43	3.12 3.94	4.92 2.31	3.75 4.0	4.56 3.22	4.08 3.12	2.5 3.5	5.0 3.5	5.0 1.0	4.0 3.0
Imitating	4.43 2.14	2.69 3.88	4.69 2.38	4.0 3.75	4.11 2.67	4.24 2.92	2.5 3.0	5.0 3.5	5.0 1.0	4.0 2.0

Table 4. Item-by-item comparison averages; **teachers** only, **sentences** only (conventions as in Table 2)

Language	**DA**	**EN**	**FA**	**FR**	**IS**	**IE**	**IT**	**ZH**	**SP**	**SW**
(#raters)	(7)	(16)	(13)	(4)	(9)	(25)	(2)	(2)	(1)	(1)
(#items)	(14)	(8)	(20)	(22)	(15)	(39)	(23)	(17)	(16)	(14)
Human acceptable	98.0	100.0	97.3	93.2	94.8	97.9	84.8	100.0	100.0	100.0
TTS acceptable	41.8	92.2	55.0	97.7	81.5	98.5	76.1	58.8	0.0	100.0
Human better	81.6	31.2	78.1	14.8	60.0	33.2	30.4	47.1	100.0	85.7
TTS better	3.1	32.8	4.2	28.4	15.6	11.8	30.4	2.9	0.0	0.0
(same)	(15.3)	(35.9)	(17.7)	(56.8)	(24.4)	(55.0)	(39.1)	(50.0)	(0.0)	(14.3)

4. Discussion and conclusions

This is a preliminary study in a rapidly evolving field, mostly using only one text per language, with genres ranging from simple children's stories to literary novels. The human voices were a mixture of male and female ranging from experienced teachers to a twelve-year-old child, while all but one of the TTS voices were young females.

With the above caveats, human audio was more often preferred than TTS, but this was by no means always the case; the gap was surprisingly close. Some TTS engines are better than others: the English, Irish, and Italian speech engines used clearly outperform the Danish and Farsi ones. TTS engines did very well on pronunciation (high scores in the 'Words' rows), but less well on sentence-level phenomena such as prosody, coarticulation processes, speed, etc. (lower scores in the 'Sentences' rows). Teachers rated TTS more highly than native speakers did (comparing Table 1 and Table 2 with Table 3 and Table 4). Non-native speakers and non-teachers rated TTS even more highly (see supplementary materials).

We are planning an extended study using a larger sample of texts and voices.

5. Supplementary materials

Relevant LARA texts, data collection form, and full results: https://www.issco.unige.ch/en/research/projects/callector/EUROCALL_2021_data.html

6. Acknowledgments

We would like to thank Hossein, Matilde, and Rebe for recording the French, Italian, and Spanish audio respectively.

References

Akhlaghi, E., Bédi, B., Bektas, F., Berthelsen, H., Butterweck, M., Chua, C., Cucchiarini, C., Eryigit, G., Gerlach, J., Habibi, H., Ní Chiaráin, N., Rayner, M., Steingrímsson, S., & Strik, H. (2020). Constructing multimodal language learner texts using LARA: experiences with nine languages. *Proceedings of the 12th Language Resources and Evaluation Conference*. https://aclanthology.org/2020.lrec-1.40