Using the LARA Little Prince to compare human and TTS audio quality

Akhlaghi, Elham; Auðunardóttir, Ingibjörg; Bączkowska, Anna; Bédi, Branislav; Beedar, Hakeem; Berthelsen, Harald; Chua, Cathy; Cucchiarini, Catia; Habibi, Hanieh; Horváthová, Ivana; Ikeda, Junta; Maizonniaux, Christèle; Chiaráin, Neasa Ní; Raheb,&nbspChadi [**and 4 more**]

Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2967–2975
Marseille, 20-25 June 2022

Using the LARA Little Prince to compare human and TTS audio quality

Elham Akhlaghi[1], Ingibjörg Iða Auðunardóttir[2], Anna Bączkowska[3], Branislav Bédi[4], Hakeem Beedar[5], Harald Berthelsen[6], Cathy Chua[7], Catia Cucchiarini[8], Hanieh Habibi[9], Ivana Horváthová[10], Junta Ikeda[7], Christèle Maizonniaux[11], Neasa Ní Chiaráin[6], Chadi Raheb[12], Manny Rayner[9], John Sloan[6,9], Nikos Tsourakis[9], Chunlin Yao[13]

[1]Ferdowsi University of Mashhad, Iran; [2]University of Iceland, Iceland;
[3]University of Gdansk, Gdansk, Poland; [4]The Árni Magnússon Institute for Icelandic Studies, Iceland;
[5]University of Adelaide, Australia;[6]Trinity College, Dublin, Ireland; [7]Independent scholar;
[8]Centre for Language and Speech Technology (CLST), Radboud University Nijmegen, The Netherlands;
[9]FTI/TIM, University of Geneva, Switzerland; [10]Constantine the Philosopher University, Nitra, Slovakia;
[11]Flinders University, Adelaide, Australia; [12]University of Guilan, Rasht, Iran
[13]Tianjin Chengjian University, Tianjin, China
elhamakhlaghi80@gmail.com, iia2@hi.is, anna.baczkowska@ug.edu.pl, branislav.bedi@arnastofnun.is,
hbeedar@hotmail.com.au, berthelh@tcd.ie, cathyc@pioneerbooks.com.au,
c.cucchiarini@let.ru.nl, hanieh.habibi@unige.ch, ihorvathova@ukf.sk, ikedaj_91@hotmail.com,
christele.maizonniaux@flinders.edu.au, Neasa.NiChiarain@tcd.ie,chadi.raheb@gmail.com,
Emmanuel.Rayner@unige.ch, john.sloan.1@ucdconnect.ie,
Nikolaos.Tsourakis@unige.ch, yao_chunlin@126.com

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 second and foreign (L2) learning through reading. An important question is how to create good quality audio, which can be done either through human recording or by a Text-To-Speech (TTS) engine. We may reasonably expect TTS to be quicker and easier, but human to be of higher quality. Here, we report a study using the open source LARA platform and ten languages. Samples of audio totalling about five minutes, representing the same four passages taken from LARA versions of Saint-Exupèry's *Le petit prince*, were provided for each language in both human and TTS form; the passages were chosen to instantiate the 2×2 cross product of the conditions {dialogue, not-dialogue} and {humour, not-humour}. 251 subjects used a web form to compare human and TTS versions of each item and rate the voices as a whole. For the three languages where TTS did best, English, French and Irish, the evidence from this study and the previous one it extended suggest that TTS audio is now pedagogically adequate and roughly comparable with a non-professional human voice in terms of exemplifying correct pronunciation and prosody. It was however still judged substantially less natural and less pleasant to listen to. No clear evidence was found to support the hypothesis that dialogue and humour pose special problems for TTS. All data and software will be made freely available.

Keywords: TTS, evaluation, multimodality, reading, emotion

1. Introduction and overview

In this paper, our central goal is to investigate the feasibility of using modern TTS technology in CALL applications. The essential parameters of the problem are straightforward. People who are trying to improve their abilities in L2 languages using a CALL application will often need L2 audio to train their listening skills. This audio can be provided by human voices or TTS engines. TTS makes it possible to produce audio quickly and cheaply and also offers various technical advantages, in particular the ability to control tempo (reading speed) and add timing information that supports synchronised visual highlighting of words. A good human voice will, however, offer higher quality. In particular, human readers understand the text at a deeper level, and may be able to use their voices to convey emotional and dramatic aspects. The aim in this article is to gain insight into how to balance these competing dimensions; we want to know when we can expect TTS quality to be good enough to be useful for L2 learning using CALL applications, and which criteria should guide this choice. A use case that particularly interests us is the teacher who wishes to create a piece of multimodal L2 material for their classroom. The speed and convenience of TTS compared to human recording can mean the difference between being able to do this and deciding that there is insufficient time. But being able to create a TTS-based resource quickly is irrelevant if the quality is too low for it to be pedagogically adequate.

People who are only familiar with previous-generation TTS may not think the question is interesting, since the quality of TTS-produced audio was until recently quite low. However, quality has improved dramatically over the last decade as Deep Neural Net (DNN) methods have become widespread. Modern TTS engines are trained from samples of recorded human audio using DNN-based toolkits. As well as improving the quality of the audio, this has also made it much simpler to create TTS voices. A high-quality TTS voice for a new language can now be created from just a few hours of recorded audio using a freely available platform.

For TTS, the first hurdle is intelligibility: is the generated TTS audio clear enough to be readily comprehensible? For the best systems, this hurdle was cleared some time ago

(King, 2014). But even if a TTS voice is comprehensible, it may still sound very artificial. The new challenges are concerned with less tangible issues (Andersson et al., 2012; Georgila et al., 2012; Pincus et al., 2015): does the voice sound natural, is it pleasant to listen to for extended periods, is it evocative?

Here, we are most concerned with pedagogical adequacy. Even if TTS-generated audio is not as good as the best human audio, is it good enough that it will instil productive habits in students who use it as a model for learning to understand and produce spoken language; more specifically, how does it compare to audio that has been recorded by the non-professional voices which typically are available in practice? This specific question does not seem to have been directly studied, but related ones have received considerable attention. An important recent example is the study presented in (Cambre et al., 2020), which compared eighteen TTS voices and three human voices, using a large number of evaluators: one of the key conclusions was that the best TTS engines are now preferred to at least some good non-professional voices. We will return to this in the final section.

To make the discussion concrete, we focus on the production of annotated multimedia texts, by which we mean texts containing integrated help, including audio and typically also translations and other types of information. This idea has become increasingly popular; high-profile examples include LingQ[1] and Learning With Texts[2]. We report experiments carried out using LARA[3], an open source platform for producing annotated multimedia texts, and contrast human-audio and TTS-audio versions of LARA texts. In most cases, the original text had human-recorded audio and a parallel version was created using a TTS engine integrated into the LARA platform. Human audio was created by non-professional voice talents, either volunteers or students working for minimal financial compensation. TTS audio was produced by the best TTS engine available for the language in question. We collected data comparing human and TTS audio quality using an open online questionnaire and collated the results in tabular form.

The work reported in this paper is a direct continuation of a previous study carried out using the same methodology (Akhlaghi et al., 2021), which reported results from an experiment with LARA texts in ten languages and 130 evaluators. Unexpectedly, the balance between human audio and TTS audio was quite close, with the native speaker evaluators in three of the ten languages — French, Irish and Swedish — on average rating the TTS version as equal to or better than the human version on full sentences. However, since different texts were used in each language (basically, we used readily available LARA texts), it was often hard to determine why TTS was judged better in some cases than in others. In the present paper, we improve the methodology by leveraging the results of the LARA Little Prince project, an informal multi-site collaboration, in which volunteers are developing LARA versions of Saint-Exupèry's classic *Le petit prince* (de Saint-Exupéry, 1945) in many languages. We use the same four short passages from *Le petit prince* for ten languages. Since a reasonable hypothesis, based on analysis of the data from the first experiment, was that dialogue and humour might both pose special problems for TTS, we chose the passages so that they instantiated the 2×2 cross product of the conditions {dialogue, not-dialogue} and {humour, not-humour}. We obtained results from 251 evaluators and collated them using slightly extended versions of the scripts used in the last study.

The rest of the paper is structured as follows. In §2., we describe the experiment in detail. §3. presents the results, and §4. discusses what we can learn from them. §5. concludes and suggests further directions.

2. Design of the experiment

We describe in turn the LARA platform, the texts, the TTS engines and human voices, and the web questionnaire used to collect the data.

2.1. The LARA platform

The experiments were performed using LARA (Akhlaghi et al., 2019; Akhlaghi et al., 2020), an open source learning-by-reading platform under development by an international consortium since 2018. LARA supports easy construction of annotated multimodal texts using open source tools which can either be invoked from the command-line or, more commonly, through an online portal. The platform provides support for crowdsourced collaborative work.

LARA texts typically include integrated audio, translations, and an automatically generated concordance. A screenshot of a page from one of the LARA texts used in our experiment is shown in Figure 1.

2.2. Texts used: the LARA Little Prince project

Since Q2 2020, volunteers in several countries have been creating LARA versions of *Le petit prince*. Choosing this book as a common text makes good sense; it is an extremely popular low intermediate reader suitable for both children and adults, which is out of copyright and translated into several hundred languages. As of January 2022, full LARA versions exist for French, English and Italian, there are usable partial versions for Polish, Icelandic and Japanese, and at least a few chapters have been completed for Irish, Farsi, Mandarin and Slovak. Links to many of these texts are posted on the LARA examples page[4]. This work is described in more detail elsewhere (Akhlaghi et al., to appear 2022).

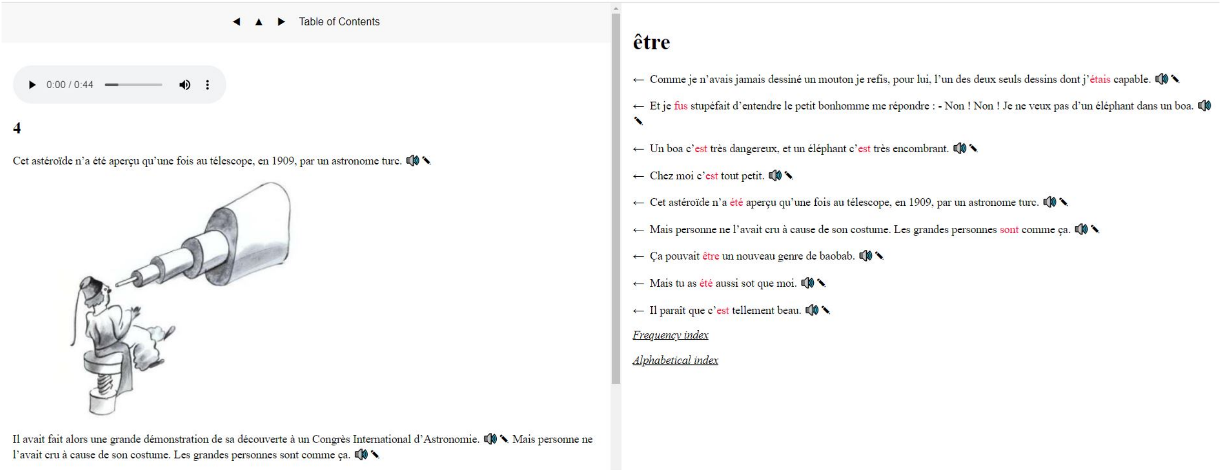
For the experiment described here, we selected passages exemplifying the distinction between dialogues and narratives and humorous and non-humorous tone. "Dialogue" was defined as requiring an explicit verbal exchange between two interlocutors, but "humorous" was entirely subjective. The passages selected were the following, instantiating the four possible combinations of {dialogue, not-dialogue} × {humour, not-humour}:

[1] https://www.lingq.com/

[2] https://sourceforge.net/projects/lwt/

[3] https://www.unige.ch/callector/lara

[4] https://www.unige.ch/callector/lara-content



4

Cet astéroïde n'a été aperçu qu'une fois au télescope, en 1909, par un astronome turc. 🔊 ✎

Il avait fait alors une grande démonstration de sa découverte à un Congrès International d'Astronomie. 🔊 ✎ Mais personne ne l'avait cru à cause de son costume. Les grandes personnes sont comme ça. 🔊 ✎

être

← Comme je n'avais jamais dessiné un mouton je refis, pour lui, l'un des deux seuls dessins dont j'étais capable. 🔊 ✎

← Et je fus stupéfait d'entendre le petit bonhomme me répondre : - Non ! Non ! Je ne veux pas d'un éléphant dans un boa. 🔊 ✎

← Un boa c'est très dangereux, et un éléphant c'est très encombrant. 🔊 ✎

← Chez moi c'est tout petit. 🔊 ✎

← Cet astéroïde n'a été aperçu qu'une fois au télescope, en 1909, par un astronome turc. 🔊 ✎

← Mais personne ne l'avait cru à cause de son costume. Les grandes personnes sont comme ça. 🔊 ✎

← Ça pouvait être un nouveau genre de baobab. 🔊 ✎

← Mais tu as été aussi sot que moi. 🔊 ✎

← Il paraît que c'est tellement beau. 🔊 ✎

Frequency index

Alphabetical index

Figure 1: Screenshot of the French LARA text (online here) used in the current study. The control at the top lets the student play the whole page. Clicking on a loudspeaker icon plays audio for the preceding sentence; clicking on a word plays audio for the word and also shows a concordance on the right (here shown for *être*, dictionary form of *été*, "been", in the first line). Hovering over a pencil icon shows a translation for the preceding sentence.

Dialogue, humour: the narrator meets the Little Prince in the desert, ch 2.

Not-dialogue, humour: the discovery of asteroid B651 by the Turkish astronomer, beginning of ch 4.

Not-dialogue, not-humour: the appearance of the Flower, beginning of ch 8.

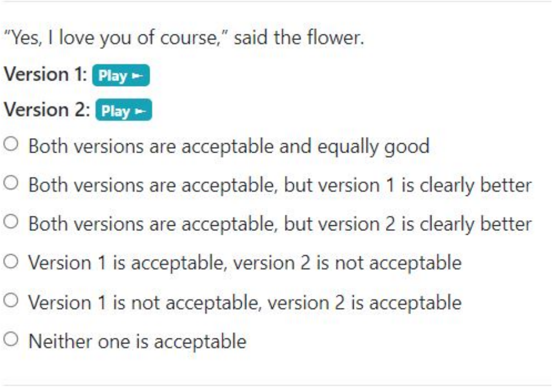
Dialogue, not-humour: the Little Prince leaves the Flower, ch 9.

In cases where the relevant human recordings did not already exist, the voice talents recorded a long enough passage, typically around two minutes, that the character of the passage was clearly established. In a second phase, the passages were carefully shortened so that the total audio which evaluators listened to (human and audio combined) was about five minutes, the exact amount varying between languages.

The passages were read by a variety of voices. French, Italian and English were read by young preteens/teens (M, F and F respectively), paid at the local babysitting rate; Icelandic by a female student in her 20s; Farsi, Irish and Slovak by female academics in their 30s; Polish by four different students (F, M, F and F respectively) in their 20s; Japanese by a graduate student and a young professional, both males in their 20s; and Mandarin by a male academic in his early 40s.

TTS was produced by the best TTS voice available to us for each language. In practice, this was ReadSpeaker[5] for English, French, Icelandic and Italian; Google Cloud TTS[6] for Japanese, Mandarin, Polish and Slovak; ABAIR[7] for Irish; and Nuance Vocalizer[8] for Farsi. All of these except Vocalizer are integrated into LARA, making it very easy to create the TTS versions.

Links to the online LARA documents are given in Table 1. The different number of segments in each version is due to the fact that the creators decided independently how to convert the original text into LARA form.



"Yes, I love you of course," said the flower.

Version 1: Play ►

Version 2: Play ►

○ Both versions are acceptable and equally good

○ Both versions are acceptable, but version 1 is clearly better

○ Both versions are acceptable, but version 2 is clearly better

○ Version 1 is acceptable, version 2 is not acceptable

○ Version 1 is not acceptable, version 2 is acceptable

○ Neither one is acceptable

Figure 2: Typical extract from "item-by-item" part of the web form.

2.3. Web form

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. Subjects were not told that one voice would be human and one TTS.

In the "demographic data" section, subjects chose the language they would use, specified their level of expertise (choice of "Native", "Near-native", "Advanced", "Intermediate" and "Beginner"), gave their gender, year of birth and level of education, and specified whether or not they had experience in teaching the language in question, a hearing

[5] https://www.readspeaker.com/

[6] https://cloud.google.com/text-to-speech

[7] abair.ie

[8] https://www.nuance.com/omni-channel-customer-engagement/voice-and-ivr/text-to-speech/vocalizer.html

Table 1: LARA texts used for experiments. “Lang” = language;“#Seg” = number of segments; “Links” = links to online material; “Hum/o” = human audio version (original); “TTS/o” = TTS audio version (original); “Hum/s” = human audio version (shortened); “TTS/s” = TTS audio version (shortened).

Lang	#Seg	Links			
		Hum/o	TTS/o	Hum/s	TTS/s
English	28	☞	☞	☞	☞
Farsi	28	☞	☞	☞	☞
French	34	☞	☞	☞	☞
Icelandic	37	☞	☞	☞	☞
Irish	40	☞	☞	☞	☞
Italian	31	☞	☞	☞	☞
Japanese	39	☞	☞	☞	☞
Mandarin	37	☞	☞	☞	☞
Polish	29	☞	☞	☞	☞
Slovak	33	☞	☞	☞	☞

impairment, or a reading impairment.

The main body of the questionnaire consisted of the “item-by-item comparisons” section, illustrated in Figure 2. The human and TTS versions of the audio were presented as “Version 1” and “Version 2”, with the order randomly selected but constant across the form, so that either Version 1 was always human, or Version 1 was always TTS. Subjects were instructed to listen to both versions and then choose the answer which they thought fitted best. The intention was to determine (a) how often each version was considered acceptable, and (b) how often one version was considered clearly better than the other. Subjects were told to interpret “acceptable” as meaning “acceptable for the purpose of learning the language in question”.

In the “overall impressions” section, subjects rated each voice, identified as “Version 1” and “Version 2”, on a five-point Likert scale, for quality of individual words, quality of whole sentences, speed, naturalness, pleasantness, suitability for teaching, suitability for imitating, and a freeform response. The notion of “quality” was left deliberately vague. We considered more fine-grained questions, for example asking for quality of pronunciation, prosody, etc, but decided against them on the basis of previous experience; it seemed likely that many subjects on an open questionnaire would be too uncertain about the meanings of even mildly technical terms.

A link to an online document with screenshots showing all the screens in the survey tool can be found in Table 2.

3. Results

We obtained questionnaire data from 251 anonymous subjects, of whom 145 self-identified as “native” or “near-native” and 124 as “having teaching experience” in the relevant language. We collated this material into tabular form using a slightly modified version of the Python script from (Akhlaghi et al., 2021). There are both quantitative and qualitative results. Full data in both raw and tabular form is posted on the web as detailed in Table 2.

Table 2: Links to full data posted on web.

Resource	Link
Item-by-item comparisons (tabular)	☞
Likert-scale ratings for voices (tabular)	☞
Freeform comments on voices (text)	☞
Raw data in JSON form	☞
Screenshots from survey tool	☞

3.1. Quantitative results

The main quantitative results are presented in Figures 3 and 4. In Figure 3, we show item-by-item comparison and Likert-scale averages for the “native/near-native” and “having teaching experience” subsets. The item-by-item comparison sections show average scores for each language in the categories “human audio judged acceptable”, “TTS audio judged acceptable”, “human audio judged better than TTS audio”, “TTS audio judged better than human audio” and “human and TTS audio judged as equal”. The Likert-scale sections contrast average score (1–5 scale, high = good) for human voice versus average score for TTS voice for each language on the questions “Individual words were correctly pronounced”, “Each sentence as a whole was correctly pronounced”, “Speed of speech was appropriate.”, “The voice sounded natural”, “The voice was pleasant to listen to”, “The voice would be acceptable for teaching purposes” and “I would recommend learners to use this voice as a model for imitating”. Colours are used to highlight cells where TTS is equal to or better than human (yellow) or close (orange).

Figure 4 presents item-by-item comparison data, using native/near-native evaluators only, that contrasts the “dialogue” passages (extracts from Chapters 2 and 9) against the “non-dialogue” passages (extracts from Chapters 4 and 8) and the “humour” passages (extracts from Chapters 2 and 4) against the “non-humour” passages (extracts from Chapters 8 and 9). The conventions used are the same as in Figure 3.

3.2. Qualitative results

We also collected qualitative results in the shape of freeform comments from the evaluators. Table 3 gives the number of comments for each language.

4. Discussion

The study described here represents a clear improvement on the one reported in (Akhlaghi et al., 2021). All languages used the same text, making comparisons between languages meaningful, and the text chosen, *Le petit prince*, was one which could plausibly be used as learning material for low intermediate readers. Nearly twice as much data was collected (251 subjects against 130 subjects). The question which most interests us is one that is both practically and pedagogically relevant: when constructing a CALL application like a LARA text, which requires audio that learners

(a) Item-by-item comparisons (native/near-native speaker evaluators)

language	English	Farsi	French	Icelandic	Irish	Italian	Japanese	Mandarin	Polish	Slovak
(#raters)	*(11)*	*(10)*	*(13)*	*(34)*	*(8)*	*(5)*	*(6)*	*(9)*	*(22)*	*(27)*
(#items)	*(28)*	*(28)*	*(34)*	*(37)*	*(40)*	*(31)*	*(39)*	*(37)*	*(29)*	*(33)*
(#annotations)	*(308)*	*(280)*	*(442)*	*(1258)*	*(320)*	*(155)*	*(234)*	*(333)*	*(638)*	*(891)*
human_acceptable	92.2	98.9	90.7	98.5	100.0	99.4	99.1	100.0	93.9	98.1
tts_acceptable	95.8	49.3	97.1	75.4	98.4	83.9	63.7	90.1	95.1	84.7
human_better	25.6	83.6	16.5	63.4	41.9	45.8	62.0	64.9	31.8	47.9
tts_better	41.2	0.7	21.0	2.8	1.2	8.4	2.6	2.1	32.4	7.4
(same)	(33.1)	(15.7)	(62.4)	(33.8)	(56.9)	(45.8)	(35.5)	(33.0)	(35.7)	(44.7)

(b) Likert-scale ratings of voices (native/near-native speaker evaluators)

language	English	Farsi	French	Icelandic	Irish	Italian	Japanese	Mandarin	Polish	Slovak
(#raters)	*(11)*	*(10)*	*(13)*	*(34)*	*(8)*	*(5)*	*(6)*	*(9)*	*(22)*	*(27)*
words_ok	4.18/4.73	4.8/2.9	3.77/4.92	4.85/3.76	5.0/4.38	4.2/4.6	4.83/3.5	4.11/3.89	3.95/4.27	4.67/3.85
sentences_ok	3.91/4.45	4.8/3.1	4.38/4.92	4.85/3.12	4.88/4.5	4.4/3.2	5.0/2.67	3.89/3.56	4.05/3.73	4.33/3.15
speed_ok	4.0/4.64	4.6/3.5	4.38/4.69	4.56/3.38	4.75/3.5	4.6/3.4	4.5/4.17	3.67/3.67	4.09/4.05	4.59/3.74
natural	4.09/3.27	4.8/3.0	4.85/3.0	4.71/2.59	4.88/3.62	4.8/2.0	4.67/2.17	4.0/2.56	4.14/2.64	4.48/2.33
pleasant	3.36/3.64	4.4/3.4	4.46/3.54	4.71/3.21	4.88/3.62	3.8/3.0	5.0/2.33	3.89/2.78	3.36/3.32	4.41/2.93
ok_for_teaching	3.27/3.82	4.9/3.0	4.15/4.08	4.68/3.12	4.88/3.5	3.8/3.4	4.5/2.0	3.78/3.33	3.55/3.36	4.33/2.81
ok_to_imitate	2.64/3.36	4.8/2.9	3.85/3.46	4.62/2.62	4.88/3.25	3.8/3.0	4.33/1.67	3.78/2.33	3.27/3.05	4.04/2.41

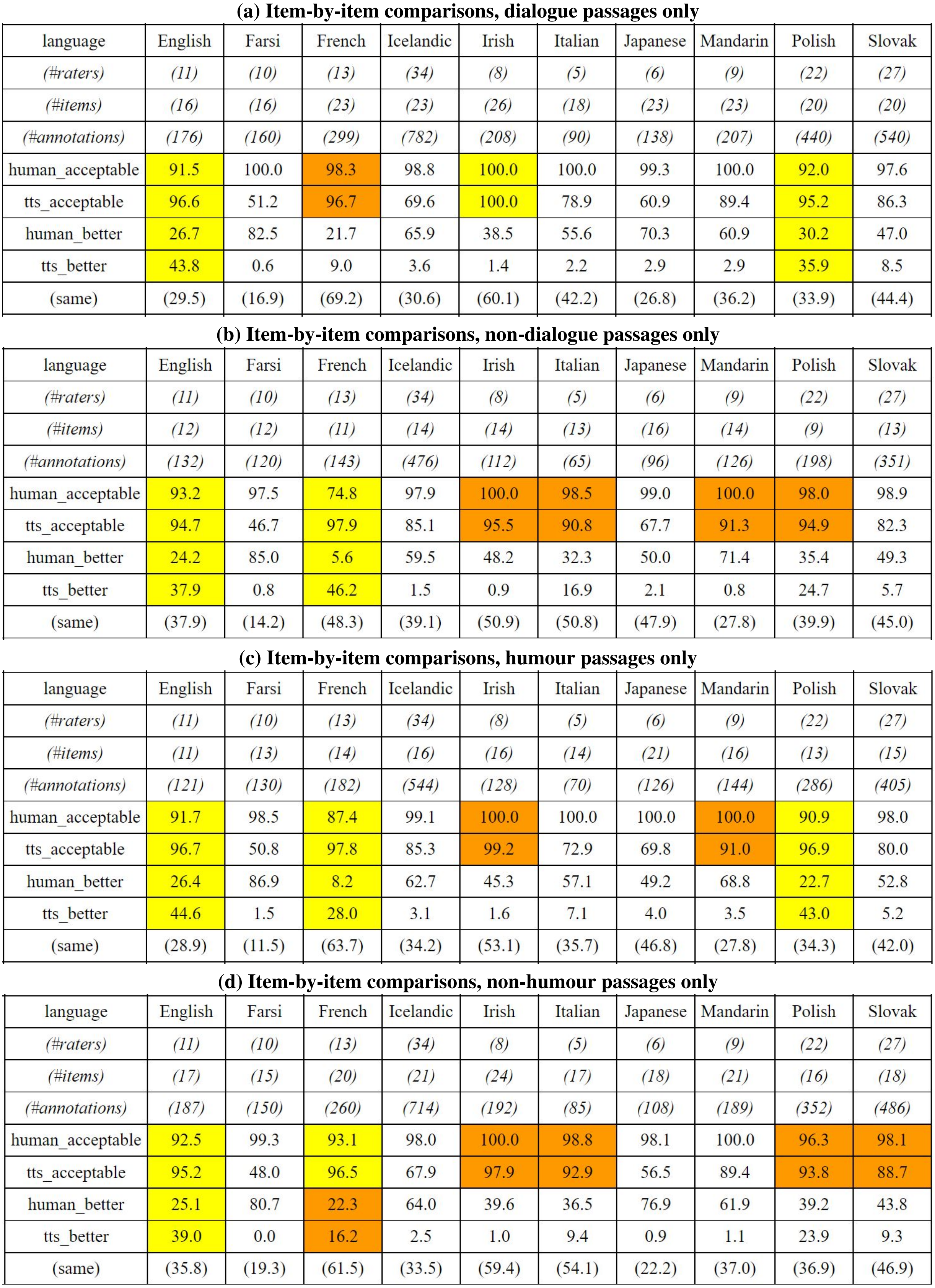
(c) Item-by-item comparisons (teacher evaluators)

language	English	Farsi	French	Icelandic	Irish	Italian	Japanese	Mandarin	Polish	Slovak
(#raters)	*(24)*	*(7)*	*(9)*	*(17)*	*(31)*	*(4)*	*(4)*	*(7)*	*(5)*	*(16)*
(#items)	*(28)*	*(28)*	*(34)*	*(37)*	*(40)*	*(31)*	*(39)*	*(37)*	*(29)*	*(33)*
(#annotations)	*(672)*	*(196)*	*(306)*	*(629)*	*(1240)*	*(124)*	*(156)*	*(259)*	*(145)*	*(528)*
human_acceptable	91.2	99.0	91.8	98.7	99.5	94.4	98.7	100.0	96.6	98.3
tts_acceptable	96.0	58.2	97.4	74.4	91.5	92.7	85.9	95.8	98.6	90.7
human_better	18.5	87.2	13.1	63.9	64.9	18.5	62.2	66.4	24.8	52.7
tts_better	44.6	1.0	17.6	1.1	1.9	25.0	3.2	3.9	51.0	13.4
(same)	(36.9)	(11.7)	(69.3)	(35.0)	(33.1)	(56.5)	(34.6)	(29.7)	(24.1)	(33.9)

(d) Likert-scale ratings of voices (teacher evaluators)

language	English	Farsi	French	Icelandic	Irish	Italian	Japanese	Mandarin	Polish	Slovak
(#raters)	*(24)*	*(7)*	*(9)*	*(17)*	*(31)*	*(4)*	*(4)*	*(7)*	*(5)*	*(16)*
words_ok	3.67/4.29	5.0/2.14	4.0/4.89	4.76/3.82	4.81/3.81	4.0/4.5	5.0/3.25	4.43/4.29	3.6/4.8	4.38/4.0
sentences_ok	3.62/4.04	5.0/2.29	4.56/4.89	4.82/2.94	4.71/3.71	4.25/3.75	5.0/3.25	4.29/3.86	4.4/4.6	4.31/3.19
speed_ok	3.46/3.96	4.71/2.86	4.44/4.67	4.82/3.53	4.84/2.35	4.75/4.25	4.5/4.75	4.14/3.71	4.4/5.0	4.5/3.56
natural	3.92/2.92	5.0/2.29	4.78/2.89	4.76/2.65	4.65/2.55	4.75/2.75	5.0/2.25	4.57/2.57	4.0/3.6	4.56/2.12
pleasant	3.33/3.29	4.71/2.43	4.56/3.0	4.65/3.29	4.81/2.35	4.0/3.0	5.0/2.25	4.43/2.71	3.0/4.0	4.38/2.88
ok_for_teaching	2.96/3.62	5.0/2.0	4.22/3.78	4.65/3.18	4.9/2.58	3.75/3.5	4.25/2.5	4.0/3.57	2.6/4.2	4.44/2.94
ok_to_imitate	2.79/3.42	5.0/1.86	4.0/2.89	4.59/2.53	4.74/2.29	3.25/3.75	4.0/1.5	3.86/2.57	2.8/4.0	4.25/2.38

Figure 3: Results for native/near-native speaker and teacher evaluators. Item-by-item comparisons: percentages, yellow = “TTS equal or better than human”, orange = “TTS within 10% of human”. Likert scale ratings: human rating/TTS rating, yellow = “TTS equal or better than human”, orange = “TTS within 0.5 of human”.



(a) Item-by-item comparisons, dialogue passages only

language	English	Farsi	French	Icelandic	Irish	Italian	Japanese	Mandarin	Polish	Slovak
(#raters)	*(11)*	*(10)*	*(13)*	*(34)*	*(8)*	*(5)*	*(6)*	*(9)*	*(22)*	*(27)*
(#items)	*(16)*	*(16)*	*(23)*	*(23)*	*(26)*	*(18)*	*(23)*	*(23)*	*(20)*	*(20)*
(#annotations)	*(176)*	*(160)*	*(299)*	*(782)*	*(208)*	*(90)*	*(138)*	*(207)*	*(440)*	*(540)*
human_acceptable	91.5	100.0	98.3	98.8	100.0	100.0	99.3	100.0	92.0	97.6
tts_acceptable	96.6	51.2	96.7	69.6	100.0	78.9	60.9	89.4	95.2	86.3
human_better	26.7	82.5	21.7	65.9	38.5	55.6	70.3	60.9	30.2	47.0
tts_better	43.8	0.6	9.0	3.6	1.4	2.2	2.9	2.9	35.9	8.5
(same)	(29.5)	(16.9)	(69.2)	(30.6)	(60.1)	(42.2)	(26.8)	(36.2)	(33.9)	(44.4)

(b) Item-by-item comparisons, non-dialogue passages only

language	English	Farsi	French	Icelandic	Irish	Italian	Japanese	Mandarin	Polish	Slovak
(#raters)	*(11)*	*(10)*	*(13)*	*(34)*	*(8)*	*(5)*	*(6)*	*(9)*	*(22)*	*(27)*
(#items)	*(12)*	*(12)*	*(11)*	*(14)*	*(14)*	*(13)*	*(16)*	*(14)*	*(9)*	*(13)*
(#annotations)	*(132)*	*(120)*	*(143)*	*(476)*	*(112)*	*(65)*	*(96)*	*(126)*	*(198)*	*(351)*
human_acceptable	93.2	97.5	74.8	97.9	100.0	98.5	99.0	100.0	98.0	98.9
tts_acceptable	94.7	46.7	97.9	85.1	95.5	90.8	67.7	91.3	94.9	82.3
human_better	24.2	85.0	5.6	59.5	48.2	32.3	50.0	71.4	35.4	49.3
tts_better	37.9	0.8	46.2	1.5	0.9	16.9	2.1	0.8	24.7	5.7
(same)	(37.9)	(14.2)	(48.3)	(39.1)	(50.9)	(50.8)	(47.9)	(27.8)	(39.9)	(45.0)

(c) Item-by-item comparisons, humour passages only

language	English	Farsi	French	Icelandic	Irish	Italian	Japanese	Mandarin	Polish	Slovak
(#raters)	*(11)*	*(10)*	*(13)*	*(34)*	*(8)*	*(5)*	*(6)*	*(9)*	*(22)*	*(27)*
(#items)	*(11)*	*(13)*	*(14)*	*(16)*	*(16)*	*(14)*	*(21)*	*(16)*	*(13)*	*(15)*
(#annotations)	*(121)*	*(130)*	*(182)*	*(544)*	*(128)*	*(70)*	*(126)*	*(144)*	*(286)*	*(405)*
human_acceptable	91.7	98.5	87.4	99.1	100.0	100.0	100.0	100.0	90.9	98.0
tts_acceptable	96.7	50.8	97.8	85.3	99.2	72.9	69.8	91.0	96.9	80.0
human_better	26.4	86.9	8.2	62.7	45.3	57.1	49.2	68.8	22.7	52.8
tts_better	44.6	1.5	28.0	3.1	1.6	7.1	4.0	3.5	43.0	5.2
(same)	(28.9)	(11.5)	(63.7)	(34.2)	(53.1)	(35.7)	(46.8)	(27.8)	(34.3)	(42.0)

(d) Item-by-item comparisons, non-humour passages only

language	English	Farsi	French	Icelandic	Irish	Italian	Japanese	Mandarin	Polish	Slovak
(#raters)	*(11)*	*(10)*	*(13)*	*(34)*	*(8)*	*(5)*	*(6)*	*(9)*	*(22)*	*(27)*
(#items)	*(17)*	*(15)*	*(20)*	*(21)*	*(24)*	*(17)*	*(18)*	*(21)*	*(16)*	*(18)*
(#annotations)	*(187)*	*(150)*	*(260)*	*(714)*	*(192)*	*(85)*	*(108)*	*(189)*	*(352)*	*(486)*
human_acceptable	92.5	99.3	93.1	98.0	100.0	98.8	98.1	100.0	96.3	98.1
tts_acceptable	95.2	48.0	96.5	67.9	97.9	92.9	56.5	89.4	93.8	88.7
human_better	25.1	80.7	22.3	64.0	39.6	36.5	76.9	61.9	39.2	43.8
tts_better	39.0	0.0	16.2	2.5	1.0	9.4	0.9	1.1	23.9	9.3
(same)	(35.8)	(19.3)	(61.5)	(33.5)	(59.4)	(54.1)	(22.2)	(37.0)	(36.9)	(46.9)

Figure 4: Results for **native/near-native speaker evaluators only**, contrasting dialogue versus non-dialogue passages, and humour versus non-humour passages. Percentages, yellow = "TTS equal or better than human", orange = "TTS within 10% of human"

Table 3: Numbers of freeform comments. "Lang" = language; "All" = total number of comments for language; "Nat" = number of comments from native/near-native evaluators; "Teach" = number of comments from teacher evaluators.

Lang	Human			TTS		
	All	Nat	Teach	All	Nat	Teach
English	25	7	13	26	7	15
Farsi	8	6	2	8	6	2
French	10	8	5	10	8	5
Icelandic	20	19	11	21	20	11
Irish	19	6	17	20	6	18
Italian	4	3	2	4	3	2
Japanese	6	5	2	6	5	2
Mandarin	8	7	4	8	7	4
Polish	15	9	3	18	11	4
Slovak	27	22	12	26	21	11

will listen to and in many cases imitate, is TTS-generated audio an acceptable alternative to using human audio? Substantially more data would be needed to obtain a definite answer (more texts, more TTS engines, more human voices, more evaluators), but we are now in a better position to make an informed guess.

As with the previous study, some TTS voices appear to be much better than others. Looking at the "tts_acceptable" line of Figure 3 (a), we see averages of over 95% for English, French and Irish; this is consistent with the second (TTS) part of the Likert-scale averages in the lines "words_ok" and "sentences_ok" from Figure 3 (b), which are all well over 4. The native-speaker evaluators rate the TTS as better than the human voice in English and French, and close in Irish ("human_acceptable" line of Figure 3 (a), first (human) part of the Likert-scale averages in the lines "words_ok" and "sentences_ok" from Figure 3 (b)). These results are broadly consistent with those for the same languages in (Akhlaghi et al., 2021), where English did a little worse and the other two languages were about the same. Together, the data from the two studies suggests that TTS for these three languages is close to or at the point where it can be considered adequate for this kind of task. Examination of the columns for Mandarin and Polish suggests that TTS quality for these two languages is not far behind. At the other end of the scale, evaluator data for Farsi and Japanese strongly suggests that TTS performance for the two voices used (Nuance Vocalizer for Farsi; Google Cloud TTS for Japanese) is not yet adequate for teaching purposes.

The above should be read with some important caveats. As already noted, it is necessary to evaluate on a larger sample of texts in order to reach firm conclusions. Also, the Likert-scale judgements from the "natural" and "pleasant" lines in Figure 3 (b) strongly suggest that evaluators do not like even the best TTS voices as much as the human voices they were compared against. The freeform comments show that evaluators often marked down the English and French human voices for making more careless mistakes than the TTS voice, but they still found the humans more enjoyable to listen to; TTS voices are described by a substantial minority of the evaluators as "mechanical", "dull", "monotone" or "lacking in emotion". The longer the audio passages that students are going to listen to, the more important this becomes.

In a long-term or even medium-term perspective, we think it is unwise to read much into the observed differences in scores between TTS voices for different languages, large as they are. TTS technology is improving rapidly, there are powerful and readily available generic tools for creating voices, and simply collecting better training data is often a good way to upgrade performance. It seems reasonable to expect that many or even most TTS voices will fairly soon be as good as the best ones we saw here, and the best ones will become even better. The bottom line, it seems to us, is that the best TTS voices are probably already adequate for this kind of task, and soon many TTS voices will be.

Our expectation is that high-quality human voices will remain better than TTS voices for some time. On the other hand, it is a fact that high-quality voices are not always available for teaching purposes. Not all teachers have high-quality voices, and many of them are not even native speakers; Irish is an example of a language where the *majority* of teachers are not native speakers (Ní Chiaráin and Ní Chasaide, 2020). Even when teachers are native speakers, many of them are likely to have a regional accent. Rather than being an obstacle, this might be seen as an advantage, as research shows how important it is that L2 learners be exposed to phonetic variation in their learning process (Thomson, 2018). In this connection it might be important to consider to what extent TTS voices are becoming part of the linguistic landscape. As TTS is incorporated in growing numbers of applications in our daily lives, it becomes increasingly relevant to be exposed to and to familiarise oneself with this type of speech, which might be less natural, but nonetheless an important component of lived experience. In short, there are good reasons for thinking that the choice is not either/or: if TTS technology is adequate, it may well be best for learners to listen to a mixture of human and TTS audio.

An important consideration is that it is not yet obvious how TTS engines will successfully address the problem of conveying emotion. That said, we were surprised to obtain no clear evidence from the current experiment to support the reasonable hypothesis that dialogue and humour would pose special difficulties for TTS. Comparing Figure 4 (a) against Figure 4 (b), and Figure 4 (c) against Figure 4 (d), we see substantial differences in some columns, but these differences are not consistent across the set of languages. It seems likely to us that this at least in part reflects the short passages we were forced to use in the experiment. We would ideally have used longer ones, but an initial pilot convinced us that most evaluators were not prepared to complete a questionnaire that took more than 10–15 minutes. It is also possible that a more careful methodology would have helped when selecting the passages exemplifying the humorous/non-humorous and dialogue/non-dialogue conditions, for example using a voting scheme with multiple participants. In follow-on experiments, we will do this.

5. Summary and further directions

We have presented a study in which 251 subjects, spanning 10 languages, responded to an open, anonymous web questionnaire in which they compared human and TTS versions of audio, totalling about five minutes, taken from four passages in *The Little Prince* chosen to be substantially different in character. The primary intention was to ascertain whether TTS audio was adequate for CALL applications where students would use it as a model for improving listening skills. For the three best languages, English, French and Irish, the evidence from the current study and the previous one it extended suggest that this threshold may have been reached. Some other languages were not far behind. Although TTS was judged adequate for the best languages, it was however still judged less natural and less pleasant to listen to than the human voices against which it was contrasted. No clear evidence, however, was found to support the intuitively reasonable hypothesis that dialogue and humour pose special problems for TTS.

The previously cited study by (Cambre et al., 2020), especially §§6–8, is clear-sighted about the inherent difficulties involved in drawing conclusions from this kind of experiment. Like us, they found that the best TTS voices were preferred to at least some good non-professional human voices. However, they present some important caveats. They only used one text passage; they point out that the relative importance of different evaluation criteria will vary greatly depending on the context of use; most importantly, TTS technology is developing so quickly that the detailed findings of any study of this nature are likely to be out of date by the time they are published. We entirely agree with these points, which in our case are exacerbated by the additional problem that we are working in multiple languages, not just English.

It would obviously be desirable to carry out further studies using a larger range of text samples. The open source LARA platform offers attractive possibilities for doing this. For many languages, it is already very easy to create a LARA version of a text in parallel human audio and TTS audio form, and the crowdsourcing functionality simplifies the task of creating editions with multiple different human voices, a methodological addition that has obvious attractions.

We are currently improving and packaging the audio questionnaire code so that third parties can straightforwardly create and deploy questionnaires of the kind described here; this functionality will be made available before the date of the LREC conference. If multiple groups can use these tools to carry out independent studies, it seems reasonable to hope that we will soon be at the point where a meta-study that merges the data would yield more conclusive results about the pedagogical adequacy of current TTS technology.

Acknowledgements

We would like to thank Hossein, Matilde and Avril for recording human audio in French, Italian and English respectively, and ReadSpeaker and Nuance for making their software available to us for research purposes. Work at the University of Geneva was funded by the Swiss National Science Foundation under project 100015_204503.

6. Bibliographical References

Akhlaghi, E., Bédi, B., Butterweck, M., Chua, C., Gerlach, J., Habibi, H., Ikeda, J., Rayner, M., Sestigiani, S., and Zuckermann, G. (2019). Overview of LARA: A learning and reading assistant. In *Proc. SLaTE 2019: 8th ISCA Workshop on Speech and Language Technology in Education*, pages 99–103.

Akhlaghi, E., Bédi, B., Bektaş, F., Berthelsen, H., Butterweck, B., Chua, C., Cucchiarini, C., Eryiğit, G., Gerlach, J., Habibi, H., Ní Chiaráin, N., Rayner, M., Steingrímsson, S., and Strik, H. (2020). Constructing multimodal language learner texts using LARA: experiences with nine languages. In *Proc. LREC 2020*. Paris: European Language Resources Association.

Akhlaghi, E., Bączkowska, A., Berthelsen, H., Bédi, B., Chua, C., Cuchiarini, C., Habibi, H., Horváthová, I., Hvalsøe, P., Lotz, R., Maizonniaux, C., Ní Chiaráin, N., Rayner, M., Tsourakis, N., and Yao, C. (2021). Assessing the quality of TTS audio in the LARA learning-by-reading platform. In N. Zoghlami, et al., editors, *CALL and professionalisation: short papers from EUROCALL 2021*, pages 1–5.

Akhlaghi, E., Bączkowska, A., Bédi, B., Beedar, H., Chua, C., Cucchiarini, C., Habibi, H., Horváthová, I., Maizonniaux, C., Chiaráin, N. N., Paterson, N., Raheb, C., Rayner, M., and Yao, C. (to appear 2022). Using the LARA platform to crowdsource a multilingual, multimodal Little Prince. *Beyond Philology*.

Andersson, S., Yamagishi, J., and Clark, R. A. (2012). Synthesis and evaluation of conversational characteristics in HMM-based speech synthesis. *Speech Communication*, 54(2):175–188.

Cambre, J., Colnago, J., Maddock, J., Tsai, J., and Kaye, J. (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*, pages 1–13.

de Saint-Exupéry, A. (1945). *Le petit prince: avec des aquarelles de l'auteur*. Gallimard.

Georgila, K., Black, A. W., Sagae, K., and Traum, D. (2012). Practical evaluation of human and synthesized speech for virtual human dialogue systems. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3519–3526.

King, S. (2014). Measuring a decade of progress in text-to-speech. *Loquens*, 1(1):e006–e006.

Ní Chiaráin, N. and Ní Chasaide, A. (2020). The potential of text-to-speech synthesis in computer-assisted language learning: A minority language perspective. In Alberto Andujar, editor, *Recent Tools for Computer- and Mobile-Assisted Foreign Language Learning*, chapter 7, pages 149–169. IGI Global, Hershey, PA.

Pincus, E., Georgila, K., and Traum, D. (2015). Which synthetic voice should I choose for an evocative task? In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 105–113.

Thomson, R. I. (2018). High variability [pronunciation] training (hvpt): A proven technique about which every language teacher and learner ought to know. *Journal of Second Language Pronunciation*, 4(2):208–231.