Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1047–1053
Marseille, 20-25 June 2022

Audiobook Dialogues as Training Data for Conversational Style Synthetic Voices

Liisi Piits, Hille Pajupuu, Heete Sahkai, Rene Altrov, Liis Ermus, Kairi Tamuri, Indrek Hein, Meelis Mihkla, Indrek Kiissel, Egert Männisalu, Kristjan Suluste, Jaan Pajupuu
Department of Speech Research and Technology, Institute of the Estonian Language, Tallinn, Estonia
{liisi.piits}@eki.ee

Abstract

Synthetic voices are increasingly used in applications that require a conversational speaking style, raising the question as to which type of training data yields the most suitable speaking style for such applications. This study compares voices trained on three corpora of equal size recorded by the same speaker: an audiobook character speech (dialogue) corpus, an audiobook narrator speech corpus, and a neutral-style sentence-based corpus. The voices were trained with three text-to-speech synthesisers: two hidden Markov model-based synthesisers and a neural synthesiser. An evaluation study tested the suitability of their speaking style for use in customer service voice chatbots. Independently of the synthesiser used, the voices trained on the character speech corpus received the lowest, and those trained on the neutral-style corpus the highest scores. However, the evaluation results may have been confounded by the greater acoustic variability, less balanced sentence length distribution, and poorer phonemic coverage of the character speech corpus, especially compared to the neutral-style corpus. Therefore, the next step will be the creation of a more uniform, balanced, and representative audiobook dialogue corpus, and the evaluation of its suitability for further conversational-style applications besides customer service chatbots.

Keywords: speech corpus, text-to-speech synthesis, speaking style, synthetic speech evaluation, chatbots, Estonian, GeMAPS

1. Introduction

Synthetic voices are increasingly used in applications that require a conversational speaking style, such as voice-enabled chatbots and automatic dubbing. This raises the question as to which type of training data yields the most suitable synthetic voices for such purposes. As a general principle, a certain amount of data that represents the desired speaking style is required in order to achieve a corresponding style in synthesis. This means that a sufficient amount of conversational data should be used to achieve a conversational style synthetic voice.

So far, relatively few attempts have been made to use genuinely spontaneous conversational training data (see e.g. Andersson et al., 2012; Székely et al., 2019; Yan et al., 2021). The acquisition of high-quality conversational data is a challenge, especially for low resource languages, and its processing is time-consuming, and requires novel technical solutions. Moreover, genuine spontaneous data may not be suitable for certain conversational style applications. For instance, Andersson et al. (2012) found that conversational utterances synthesised with voices trained on spontaneous data (using techniques based on hidden Markov models (HMM)) were perceived as more natural and suitable only when they contained discourse markers and filled pauses. In cases of more fluent conversational style utterances, listeners preferred voices that were based on standard read-aloud training data.

A further question is how human the voice of an application can sound without creating discomfort and eeriness that the user may experience from interacting with a virtual communicator that is too human; this is known as the uncanny valley effect (Mori, 1970; Mori et al., 2012; see also Moore, 2012). According to Ciechanowski et al. (2019), simpler systems without an animated avatar provoked fewer uncanny effects and negative reactions than those that imperfectly imitated a human. Moore (2017) found that it is beneficial for robots to have a distinctly artificial voice, so that the user understands they are interacting with a fully automated system. For example, in contrast to a humanlike voice, there was no perceived need to tell a telephone-based travel planning service with a robotic voice the reasons for the travel. Yet many studies have shown that the more similar an artificial voice is to a human voice, the less eerie and more likeable the voice is to users (e.g. Kühne et al., 2020).

The goal of the present study was to test audiobook dialogues as a potential novel source for training conversational style synthetic voices. A major advantage of audiobook dialogues compared to spontaneous data is that they are easy to acquire and process, and have high studio-level recording quality. As an existing resource, they are also more affordable than the recording of dedicated training data in a studio. However, copyright issues may need to be solved.

From the point of view of the speaking style, audiobook dialogues could provide suitable training data for applications that require a fluent and not too human-like conversational style, as they imitate conversational speech without the characteristic disfluencies of spontaneous conversation. We used customer service voice chatbots without an animated avatar as an example of this type of conversational style application to test the suitability of audiobook dialogues as training data. Customer service voice applications are expected to render relatively standard written-style text in a formal/polite, fluent, and intelligible conversational style. The voice should also engender trust, and be characteristic and context appropriate (see e.g. Cambre and Kulkarni, 2019; Torre et al., 2018; Troshani et al., 2021). It is therefore plausible that synthetic voices trained on audiobook dialogues are suitable for such applications.

In order to test the suitability of audiobook dialogues as training data for conversational style synthetic voices, we compared audiobook dialogues with two further types of read-aloud training data recorded by the same speaker: a neutral-style sentence-based text-to-speech synthesis (TTS) corpus, and a narrator text corpus based on the same audiobooks from which the dialogues were extracted (see section 2.1 and 2.2 for details). In order to test whether the evaluations remain stable across different TTS techniques, we used three different techniques to train the synthetic voices that were tested (section 2.3). The resulting nine synthetic voices were used to synthesise a set of real

customer service chatbot speech turns that were subjected to evaluation as to the suitability of their speaking style (section 2.4).

We tested the following hypothesis: when judging the appropriateness of the speaking style of synthesised customer service chatbot speech turns, listeners prefer voices trained on audiobook dialogues to voices trained on a neutral TTS corpus and on audiobook narrator speech, independently of the TTS technique used.

2. Method and Procedure

2.1 Speech Corpora

For the purposes of the present study we created three speech corpora that were based on two larger pre-existing corpora: a fiction audiobook corpus and a corpus of isolated sentences produced in a neutral reading style. Section 2.1.1 will describe the two existing speech corpora used in the study. Section 2.1.2 will describe the three test corpora created for the purposes of the experiment.

2.1.1 The Existing Speech Corpora Used in the Study

Two existing speech corpora recorded by the same speaker were used in the study: a fiction audiobook sub-corpus and a neutral sentence-based corpus.

The fiction audiobook corpus[1] was compiled from book reading series recorded by Estonian Public Broadcasting. The recordings, along with the corresponding texts and licences, were obtained under a cooperation agreement between Estonian Public Broadcasting and the Institute of the Estonian Language. The recordings included a large quantity of materials from the same speaker, as the same actor had often recorded a large number of books. For the purposes of the corpus, three sub-corpora were created for which three pleasant and well-represented male voices with different timbres were selected. For some experienced actors the recordings spanned several decades; for example, for the actor PT, the earliest recordings dated from 1993. However, the earliest recordings were excluded from the corpus as they were not accompanied by electronic texts, the recording quality varied and, first and foremost, the human voice changes over time. The speech was segmented using WebMaus tools (Kisler et al., 2017) in order to be able to divide the recordings into sentences. The resulting material was checked sentence by sentence; the text was corrected to match with the audio, and utterances with background noise or other defects were excluded. Separate annotations were created for character speech (the dialogues) and narrator speech (the remaining text), as these two styles are significantly different (Pajupuu et al., 2019). The annotation was created automatically based on punctuation marks, and then manually corrected. The character speech was further manually annotated for the gender and age of the character and the position of the introductory sentence. The entire male voice fiction corpus contains 34 hours of speech. In the present study we used the sub-corpus recorded by the actor PT. The PT sub-corpus contains readings from six books recorded between 2015 and 2020. In total, the PT sub-corpus contains six hours of narrator speech and two hours of character speech.

The second resource used in the present study is a representative corpus of isolated sentences[2] recorded by the same speaker (PT) in a neutral speaking style. The corpus was recorded using the standard script[3] created for the recording of Estonian TTS speech corpora. The script ensures the coverage of all the Estonian phonemes, phoneme transitions, and more frequent diphthongs. In addition, the script contains a selection of frequent names, numbers, expressions, and everyday phrases. The sentences of the script do not constitute a coherent text, but have been selected individually from a newspaper corpus, or created for the purposes of fulfilling the above criteria. The corpus recorded by PT contains 1,849 sentences (2.47 hours of speech). The recording was made using a dedicated recording program created at the Institute of the Estonian Language[4]. The program displays the sentences one by one on the computer screen, and records the corresponding audio directly in a separate file.

2.1.2 The Experimental Corpora Created for the Purposes of the Present Study

For the purposes of the present study, the two pre-existing corpora recorded by PT, the fiction corpus and the neutral sentence-based corpus, were used to create three experimental corpora of an equal size (99,500 characters) and with identical technical parameters (48 kHz, 16 bit, Mono, 70dB): (1) the Character Speech Corpus (CHAR) that consisted of dialogues extracted from the PT fiction sub-corpus, (2) the Narrator Speech Corpus (NARR) extracted from the PT fiction sub-corpus excluding dialogues, and (3) the Neutral Speech Corpus (NEU) that was extracted from the PT neutral sentence-based corpus. The Character Speech Corpus contained 2,063, the Narrator Speech Corpus 867, and the Neutral Speech Corpus 1,535 sentences. The number of words per sentence in the three experimental corpora is shown in Figure 1.

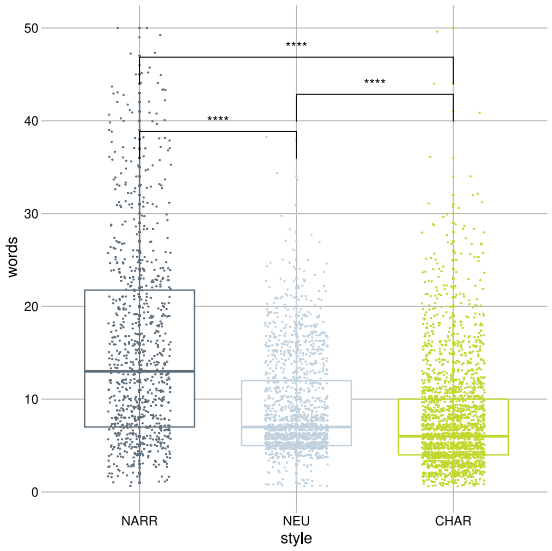

Figure 1: The number of words per sentence in the Narrator Speech (NARR), Neutral Speech (NEU) and Character Speech (CHAR) Corpora. **** $p < .0001$

[1] https://doi.org/10.15155/3-00-0000-0000-0000-08BF4L

[2] https://doi.org/10.15155/3-00-0000-0000-0000-08BF2L

[3] http://heli.eki.ee/syntees/suur_baas.doc

[4] https://koneveeb.ee/allalaadimine/salvestaja.zip

2.2 The Acoustic Description of the Experimental Corpora

We ran an acoustic analysis of the three experimental corpora in order to get an overview of the features that differentiate them. For the acoustic analysis we used the open-source toolkit openSMILE (Eyben et al., 2010, 2013). The parameters of the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) were calculated for each sentence of the corpus. These 88 parameters include statistical properties (arithmetic mean, coefficient of variation, percentiles, etc.) calculated for a set of time-varying low-level acoustic features, including frequency-related, energy-/amplitude-related, spectral, and temporal features (Eyben et al., 2016).

To identify the acoustic features that distinguish the three corpora, the Wilcoxon Test was used, and the statistically significant parameters were ordered by the test statistic (R Core Team, 2021).

The acoustic analysis showed that although the speech corpora had been built on one person's voice, they differed significantly in acoustic parameters (see Figure 2 for the most distinctive parameters; all eGeMAPS parameters for the sentences of the three corpora are available on GitHub[5]). The sentences in NEU had lower speech tempo (longer mean voiced segment length, less voiced segments per second, less loudness peaks per second), while the sentences in CHAR were marked by the fastest speech. The CHAR sentences were also the loudest (higher loudness, higher mean α ratio). Unlike CHAR and NARR sentences, NEU sentences were characterised by more uniform loudness—they featured significantly fewer rapid changes in loudness (smaller mean rising and falling slope of loudness). NEU can also be described as a corpus with a harmonic voice (higher mean HNR, flatter spectral slope), as opposed to the more rough and breathy voice of CHAR (see e.g. Gordon and Ladefoged, 2001). These properties point to the sentences of NEU having a clearer speaking style and less expressiveness, compared with audiobook corpora (see e.g. Tamuri and Mihkla 2012; Uchanski, 2005).

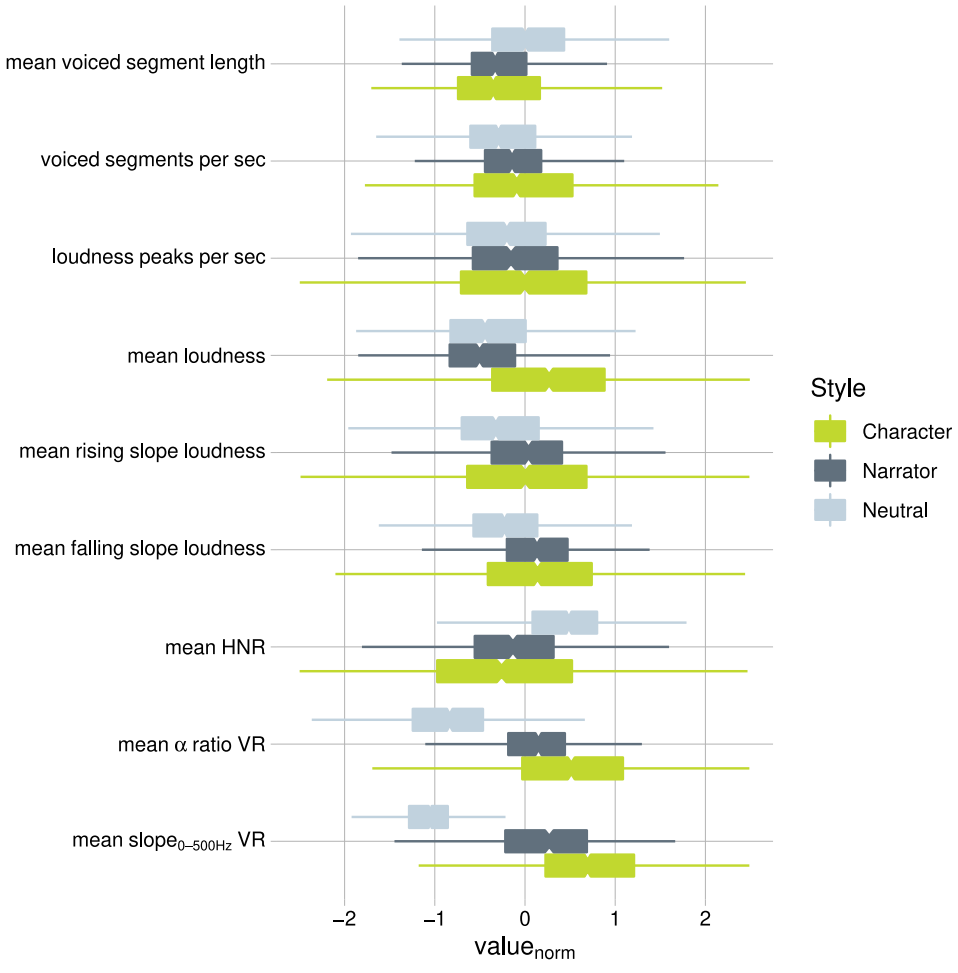

Figure 2: The acoustic parameters differentiating the sentences of the corpora. *loudness = estimate of perceived signal intensity from an auditory spectrum; rising/falling slope loudness = slope of rising/falling signal parts of loudness; HNR = harmonics-to-noise ratio; alpha ratio = ratio of the summed energy from 50–1000 Hz and 1–5 kHz; spectral slope 0–500 Hz = linear regression slope of the logarithmic power spectrum for 0–500 Hz region; VR = voiced regions*

[5] https://github.com/pajupuujh/VoiceSuitability

2.3 Text-to-Speech Synthesisers

Three existing speech synthesisers for Estonian (henceforth referred to as S1, S2, and S3) were used to train three synthetic voices on the basis of each corpus (NARR, CHAR, NEU), thus in total nine synthetic voices were trained.

S1 and S2 use HTS 2.0, a HMM-based statistical–parametric TTS technique (Zen et al., 2007). S1 uses a phoneme-based approach, and S2 uses a grapheme-based approach. This is motivated by the fact that Estonian has a relatively phonemic orthography (EKG II, 1993) with only a few distinctions that are not reflected in orthography: for example, certain quantity differences and the palatalisation of consonants (Piits and Kalvik, 2021). Estonian is thus characterised by a close correspondence between phonemes and graphemes. The phoneme-based technique (S1) uses the Estonian pronunciation rules that identify compounds, second and third quantity words, and palatalised consonants. In order to apply the pronunciation rules, the Estonian morphological analyser and disambiguator are included in the front end of the HTS-system[6]. The grapheme-based technique (S2) does not use pronunciation rules. Under this approach, language models are trained based only on grapheme sequences; during synthesis, text is directly transformed into speech without using a phonemic level. This approach uses the language independent text processing libraries of the OssianTTS[7] as the front end of the HTS system (Vainio et al., 2014).

The third synthesiser that we used **(S3)** is TransformerTTS[8], a neural network-based speech synthesiser adapted to Estonian at the University of Tartu[9]. The technique consists of three main components: a grapheme-to-phoneme converter, an acoustic model, and a vocoder. Although the acoustic model also functions with unprocessed text, the conversion of the text into phonemes using eSpeak[10] yields better results. Unfortunately, Estonian is not sufficiently supported by eSpeak, and therefore errors occur with respect to quantity and palatalisation. The acoustic model uses a transformer architecture, which considers the context of the whole sentence. This approach results in more natural sounding prosody, especially for longer sentences (Li et al., 2019). The acoustic model generates MEL spectrograms, which are converted to audio signals using a HiFiGAN vocoder[11] (Kong et al., 2020). As training vocoders requires large datasets, it is common to use pre-trained vocoders. The vocoder used in S3 was pre-trained on the VCTK corpus, which works well with previously unknown voices (Yamagishi et al., 2019).

2.4 Evaluation of the Suitability of the Synthesised Voices

A web-based listening test was carried out, where the nine synthetic voices were assessed for the suitability of their speaking style for a customer service chatbot. The rating was done by eight men and eight women (aged 31−65, M = 46.0, SD =11.1).

The listeners were presented with synthesised versions of six real customer service chatbot speech turns, each of which was synthesised with the nine synthetic voices obtained with the three corpora (NARR, CHAR, and NEU; see section 2.1) and the three synthesisers (S1, S2, and S3; see section 2.3). The duration of the test was around 15 minutes.

The sentences in the test were as follows[12]:

1. *Tere, täname sõnumi eest. [Hi, thank you for your message.]*
2. *Vastame teile esimesel võimalusel. [We will reply as soon as possible.]*
3. *Ma ei ole inimene, olen vestlusrobot Peeter. [I am not a human, I am Peeter the chatbot.]*
4. *Selleks, et saaksin õppida, salvestan meie vestluse. [I will record our conversation, so that I can learn.]*
5. *Klikake all olevat nuppu ja saan teid juhendada maksmisel. [Click the button below and I can guide you through the payment process.]*
6. *Mul on hea meel, et sain teile abiks olla! [I am glad to have been of assistance!]*

The listeners were given the following instruction: *Current customer service chatbots usually answer your questions in writing. Please imagine that instead of writing, the chatbot talks to you in Estonian. Listen to the samples and evaluate how well their speaking style would suit a chatbot.*

The listeners had to evaluate the suitability of the speaking style on a 7-point Likert scale, where 1 = not suitable at all … 7 = very suitable.

Following the execution of the listening test, all scores for each rater were normalised using the formula in (1)

$$y=(x-X)/s \qquad (1)$$

where x is the score, X is the mean of the rater's scores, and s is the standard deviation of the rater's scores. Performances with scores above zero were classified as suitable, and those with scores below zero as unsuitable.

To find out the degree of agreement among the raters (inter-rater reliability), the intra-class correlation coefficient (ICC2k) was calculated using the 'psych' package in R (Revelle, 2021).

A Welch *t*-test was used to determine whether the synthesised voice sample sets had significantly different mean scores (R Core Team, 2021).

3. Evaluation results

An excellent degree of reliability was found within rater measurements. The average measure ICC2k was .95 with a 95% confidence interval from .93 to .97 $F(53, 795) = 20$, $p < .0001$).

The results of the listening test (see Figure 3, Table 1) revealed that listeners considered all S3 voices and the S2 voice that was trained on the Neutral Speech Corpus to be suitable for a service chatbot. Voices trained on the Neutral Speech Corpus were found to be the most suitable for all synthesis techniques, except for S1 where there was no significant difference between voices trained on the Neutral

[6] https://github.com/ikiissel/synthts_et

[7] https://github.com/CSTR-Edinburgh/Ossian

[8] https://github.com/as-ideas/TransformerTTS

[9] https://github.com/TartuNLP/TransformerTTS

[10] http://espeak.sourceforge.net/

[11] https://github.com/jik876/hifi-gan

[12] The synthesised utterances used in the listening test can be found at https://github.com/pajupuujh/VoiceSuitability

and Narrator Speech Corpus. Listeners found the voices trained on the Character Speech Corpus to be the least suitable, regardless of synthesis technique.

group1	group2	stat	*df*	*p*
S3_NARR	S3_NEU	-3.04	188.76	.0030
S3_NARR	S3_CHAR	5.42	180.61	.0001
S3_NARR	S1_NARR	11.97	188.16	.0001
S3_NARR	S1_NEU	11.94	188.24	.0001
S3_NARR	S1_CHAR	17.36	181.21	.0001
S3_NARR	S2_NARR	9.38	188.46	.0001
S3_NARR	S2_NEU	7.66	184.82	.0001
S3_NARR	S2_CHAR	14.37	181.78	.0001
S3_NEU	S3_CHAR	8.23	173.99	.0001
S3_NEU	S1_NARR	15.70	189.94	.0001
S3_NEU	S1_NEU	15.66	189.95	.0001
S3_NEU	S1_CHAR	21.60	186.25	.0001
S3_NEU	S2_NARR	12.97	189.98	.0001
S3_NEU	S2_NEU	11.33	188.55	.0001
S3_NEU	S2_CHAR	18.45	186.65	.0001
S3_CHAR	S1_NARR	4.79	172.38	.0001
S3_CHAR	S1_NEU	4.76	172.59	.0001
S3_CHAR	S1_CHAR	9.05	160.72	.0001
S3_CHAR	S2_NARR	2.56	173.16	.0110
S3_CHAR	S2_NEU	0.87	165.93	**.3870**
S3_CHAR	S2_CHAR	6.51	161.48	.0001
S1_NARR	S1_NEU	-0.03	190.00	**.9750**
S1_NARR	S1_CHAR	5.02	187.12	.0001
S1_NARR	S2_NARR	-2.69	189.99	.0080
S1_NARR	S2_NEU	-4.95	189.08	.0001
S1_NARR	S2_CHAR	1.89	187.47	**.0600**
S1_NEU	S1_CHAR	5.05	187.02	.0001
S1_NEU	S2_NARR	-2.66	189.99	.0090
S1_NEU	S2_NEU	-4.91	189.02	.0001
S1_NEU	S2_CHAR	1.93	187.37	**.0560**
S1_CHAR	S2_NARR	-7.85	186.72	.0001
S1_CHAR	S2_NEU	-10.49	189.43	.0001
S1_CHAR	S2_CHAR	-3.32	189.99	.0010
S2_NARR	S2_NEU	-2.14	188.84	.0330
S2_NARR	S2_CHAR	4.73	187.09	.0001
S2_NEU	S2_CHAR	7.22	189.58	.0001

Table 1: Welch t-test for mean scores of synthesised voices

4. Discussion

We hypothesised that when judging the appropriateness of the speaking style of synthesised customer service chatbot sentences, listeners would prefer voices trained on the Character Speech Corpus (CHAR) to voices trained on the Narrator Speech Corpus (NARR) and the Neutral Speech Corpus (NEU). This hypothesis was not confirmed: the voices trained on the Character Speech Corpus (CHAR) received the lowest scores with all synthesis techniques (see Figure 3, Table 1). Somewhat surprisingly, the highest scores were received by the voices trained on the NEU corpus. The voices trained on the NARR corpus received intermediate scores with synthesisers S2 and S3, and were on a par with the voices trained on the NEU corpus with synthesiser S1. The overall level of the scores was relatively low, as the corpora used in the experiment were rather small.

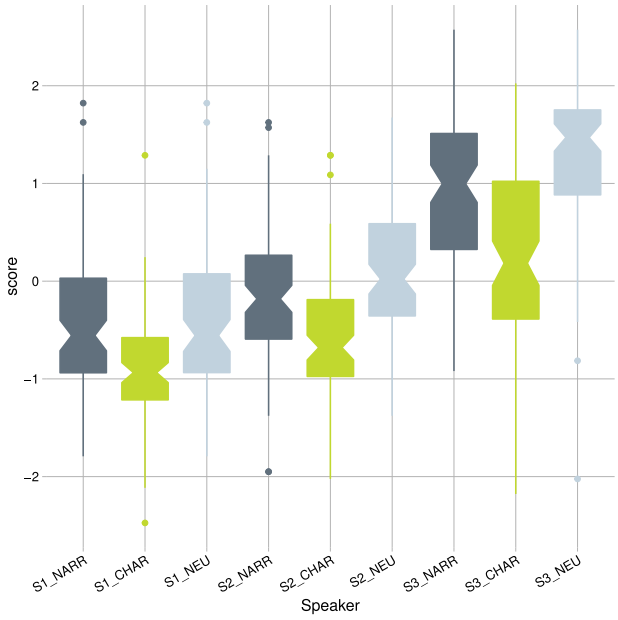

Figure 3: The normalised scores for the suitability of the speaking style of the nine customer service chatbot voices created with the different synthesisers (S1, S2, and S3) and corpora (NARR, CHAR, and NEU)

It is nevertheless promising that the voices trained on the CHAR corpus were recognisably different from the others: they received clearly different scores from the voices trained on the NARR and NEU corpora, while the voices trained on the latter two corpora did not receive different scores with S1 and differed relatively little, although significantly, with S2 and S3.

A second promising result is that the voices trained on the CHAR corpus were judged as suitable when trained with the best performing synthesiser (S3).

A confounding factor that may have influenced the ratings besides the speaking style was the quality of the synthetic voices. Several possible factors may have affected the quality of the voices trained on the CHAR corpus.

First, the CHAR corpus is characterised by a greater variability that is due to the fact that the speaker varied his voice in order to imitate different characters. The books in the corpus represented different fiction genres, and the characters that were imitated were men and women of different ages, as well as mythological creatures. The greater acoustic variability of the CHAR corpus compared to the other two corpora is also reflected in the acoustic parameters displayed in Figure 2. The CHAR corpus also contained a larger proportion of non-modal voice, as shown by lower mean HNR and steeper spectral slope (see Figure 2). A possible improvement could thus be achieved by a more careful selection of more uniform dialogues.

A second factor influencing the quality of the synthetic voices may have been the length of the utterances in the corpora. On the one hand, the length of the sentences used in the listening test was closest to the average length of the sentences in the CHAR corpus. On the other hand, the CHAR corpus contained a large number of short sentences of one to three words while sentence lengths especially in the NEU corpus were evenly distributed (see Figure 1). The quality of the voices trained on the CHAR corpus may have further suffered from the combined effect of short

sentences and the variability of F0 and timbre due to the imitation of different characters. As the language models of synthesisers are generalisations across these different imitations, such differences may result in hoarse utterance endings (this effect was especially strong with the S1 synthesiser). Thus, a further improvement could be achieved by a more even balancing of sentence lengths in dialogue corpora.

A third factor that may have affected the quality of the synthetic voices was the fact that the NEU corpus was phonemically more representative than the CHAR and NARR corpora, which may have been a great advantage as the amount of data was small. The coverage of the dialogue corpus could be improved by combining it with the neutral-style sentence-based TTS corpus.

Thus, one reason why the voices trained on the NEU corpus received the highest scores could be their highest quality due to the more representative, balanced, and homogeneous nature of the data. However, another reason could be that the NEU corpus yielded a clearer synthetic speech, as it is characterised by a slower speaking rate and hence presumably clearer articulation (see Figure 1). This could indicate that listeners expect clarity rather than a natural conversational style from a service chatbot. In fact, the customer service chatbot sentences used in the evaluation study could be taken to represent a written rather than a conversational style, thus requiring a neutral reading style rather than a conversational style. In future, voices trained on audiobook dialogues could thus be tested on other kinds of conversational applications, for example, subtitle voicing (Mihkla et al., 2014).

As far as customer service chatbot voices are concerned, and assuming that listeners do not expect them to have a conversational speaking style, audiobook narrator speech could thus provide the most suitable and affordable training data for them: it performs comparably to a neutral TTS corpus, but is an existing resource, and thus does not require special recordings to be made in a studio. Dialogues could either be excluded from the training data or selectively retained. For example, in case of a male voice, the speech of male characters could be retained, and the speech of female characters, imitated with a higher F0, excluded, so as to achieve a synthetic voice that conforms better to a male voice. Alternatively, dialogues could be used to increase the number of interrogatives in the data as these are usually rare in narrator speech.

However, a further aspect to be considered when interpreting the evaluation results is that the results are based on an imagined rather than a real situation. The results could be different in a real situation where a service chatbot with or without an animated avatar effectively answers customer's questions.

A further line of research to be considered in future is to use audiobooks recorded by other actors, so as to exclude the possibility that the results of the study were affected by the idiosyncrasies of the actor PT.

Concerning the synthesis techniques, the voices trained with the neural synthesiser S3 received significantly higher scores than the voices trained with the HMM-based synthesisers S1 and S2. Only the best performing S2 voice, S2_NEU, reached the level of the worst performing S3 voice, S3_CHAR (see Table 1). The fact that voices trained on neural networks were preferred over HMM-based voices conforms to the international experience. The finding that S3 voices, which were the most human, received the highest scores supports the study of Kühne et al. (2020) who found that the more human a synthetic voice is, the more likeable and less eerie it is.

5. Conclusion

The study compared nine synthetic voices trained on three corpora – the Character Speech Corpus, the Narrator Speech Corpus and the Neutral Speech Corpus – using three synthesisers – two HMM-based and a neural synthesiser –, evaluating the suitability of their speaking style for use in customer service voice chatbots. Independently of the synthesiser used, the voices trained on the Character Speech Corpus received the lowest, and those trained on the Neutral Speech Corpus the highest scores. However, the evaluation results may have been confounded by the greater acoustic variability, less balanced sentence length distribution, and poorer phonemic coverage of the Character Speech Corpus—especially compared to the Neutral Speech Corpus. Also, it is possible that listeners expect a customer service chatbot to speak in a slow and clearly articulated manner rather than in a distinctly conversational style. The next step will therefore be the creation of a more uniform, balanced, and representative audiobook dialogue corpus, and the evaluation of its suitability for further conversational-style applications besides customer service chatbots, for example, subtitle voicing.

6. Acknowledgements

The authors wish to express gratitude to participants of the listening test.

This work was financed by the European Union through the European Regional Development Fund (Centre of Excellence in Estonian Studies), by the National Programme for Estonian Language Technology 2018–2027, by the Estonian Research Council (project EAG144 Spontaneous speech synthesis), and by basic governmental financing of the Institute of the Estonian Language from the Estonian Ministry of Education and Research.

The authors would like to thank TalTech HPC Center for granting access to their computational resources.

7. Bibliographical References

Andersson, S., Yamagishi, J., and Clark, R. A. J. (2012). Synthesis and evaluation of conversational characteristics in HMM-based speech synthesis. *Speech Communication,* 54(2):175–188. https://doi.org/10.1016/j.specom.2011.08.001

Cambre, J. and Kulkarni, C. (2019). One voice fits all? Social implications and research challenges of designing voices for smart devices. In *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), Article 223:1–19. https://doi.org/10.1145/3359325

Ciechanowski, L., Przegalinska, A., Magnuski, M., and Gloor, P. (2019). In the shades of the uncanny valley: An experimental study of human–chatbot interaction. *Future Generation Computer Systems*, 92:539–548. https://doi.org/10.1016/j.future.2018.01.055

Eyben, F., Wöllmer, M., and Schuller, B. (2010). openSMILE: The Munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia, MM 2010*, pages 1459–146. Florence, Italy: ACM. https://doi.org/10.1145/1873951.1874246

Eyben, F., Weninger, F., Groß, F., and Schuller, B. (2013). Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM International Conference on Multimedia, MM 2013*, pages 835–838. Barcelona, Spain: ACM. https://doi.org/10.1145/2502081.2502224

Eyben, F., Scherer, K., Schuller, B., Sundberg, J., Andre, E., Busso, C., Devillers, L., Epps, J., Laukka, P., Narayanan, S., and Truong, K. (2016). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing,* 7(2):190–202. https://doi.org/10.1109/TAFFC.2015.2457417

EKG II (1993). *Eesti keele grammatika II. Süntaks.* Lisa: *Kiri* [Estonian grammar II. Syntax (Appendix: Script)], M. Erelt, T. Erelt, H. Saari, & Ü. Viks (Eds.). Eesti Keele Instituut: Tallinn.

Gordon, M. and Ladefoged, P. (2001). Phonation types: a cross-linguistic overview. *Journal of Phonetics*, 29(4):383–406. https://doi.org/10.1006/jpho.2001.0147

Kisler, T., Reichel, U., and Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, 45:326–347. https://doi.org/10.1016/j.csl.2017.01.005

Kong, J., Kim, J., and Bae, J. (2020). HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. In *The 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Vancouver, Canada. https://arxiv.org/pdf/2010.05646v2.pdf (22.11.2021)

Kühne, K., Fischer M. H., and Zhou Y. (2020). The human takes it all: Humanlike synthesized voices are perceived as less eerie and more likable. Evidence from a subjective ratings study. *Frontiers in Neurorobotics*, 14: 593732. https://doi.org/10.3389/fnbot.2020.593732

Li, N., Liu, S., Zhao, S., and Liu, M. (2019). Neural speech synthesis with transformer network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(1):6706–6713. https://doi.org/10.1609/aaai.v33i01.33016706

Mihkla, M., Hein, I., Kiissel, I., Räpp, A., Sirts, R., and Valdna, T. (2014). A system of spoken subtitles for Estonian television. In A. Utka, G. Grigonytė, J. Kapočiūtė-Dzikienė, & J. Vaičenonienė (Eds.), *Human language technologies – the Baltic perspective. Frontiers in artificial intelligence and applications,* 268, pages 19–26. IOS Press Ebooks.

Moore, R. (2012). A Bayesian explanation of the 'Uncanny Valley' effect and psychological phenomena. *Scientific Reports,* 2:864. https://doi.org/10.1038/srep00864

Moore, R. (2017). Appropriate voices for artefacts: Some key insights. In *Proceedings of the 1st International Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots (VIHAR-2017).* Skovde, Sweden.

Mori, M. (1970). Bukimi no tani (the uncanny valley). *Energy,* 7:33–35.

Mori, M., MacDorman, K. F., and Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine*, 19(2):98–100. https://doi.org/10.1109/MRA.2012.2192811

Pajupuu, H., Altrov, R., and Pajupuu, J. (2019). Towards a vividness in synthesized speech for audiobooks. *Eesti ja soome-ugri keeleteaduse ajakiri = Journal of Estonian and Finno-Ugric Linguistics*, 10(1):167−190. https://doi.org/10.12697/jeful.2019.10.1.09

Piits, L. and Kalvik, M.-L. (2021). Fonoloogiline varieerumine eesti keeles kolme nähtuse näitel. [Phonological variation in Estonian on the bases of three phenomena] *Emakeele Seltsi aastaraamat,* 66:177−201. http://dx.doi.org/10.3176/esa66.08

R Core Team (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/ (22.11.2021)

Revelle, W. (2021). *psych: Procedures for Psychological, Psychometric, and Personality Research.* Northwestern University, Evanston, Illinois. R package version 2.1.6, https://CRAN.R-project.org/package=psych (22.11.2021)

Székely, É., Henter, G., Beskow. J., and Gustafson, J. (2019). Spontaneous conversational speech synthesis from found data. In *Proceedings of Interspeech 2019,* pages 4435–4439. https://doi.org/10.21437/interspeech.2019-2836

Tamuri, K. and Mihkla, M. (2012). Emotions and speech temporal structure. In E. Meister (Ed.), *Proceedings: XXVII Fonetiikan päivät 2012 – Phonetics Symposium 2012*, pages 5−60. Tallinn: TUT Press.

Torre, I., Goslin, J., White, L., and Zanatto, D. (2018). Trust in artificial voices: A "congruency effect" of first impressions and behavioural experience. In *Proceedings of APAScience '18: Technology, Mind, and Society (TechMindSociety '18)*, pages 1–6. https://doi.org/10.1145/3183654.3183691

Troshani, I., Rao Hill, S., Sherman, C., and Arthur, D. (2021). Do we trust in AI? Role of anthropomorphism and intelligence. *Journal of Computer Information Systems*, 61(5):481–491. https://doi.org/10.1080/08874417.2020.1788473

Uchanski, R. M. (2005). Clear speech. In D. B. Pisoni and R. E. Remez (Eds.), *The Handbook of Speech Perception,* Oxford: Blackwell Publishing Ltd, pp. 207–235. https://doi.org/10.1002/9780470757024

Vainio, M., Gronroos, S.-A., Smit, P., Suni, A., and Watts, O. (2014). *Description of the final version of the new front-end.* https://simple4all.org/wp-content/uploads/2014/11/Simple4All_deliverable_D2.2.pdf (22.11.2021)

Yan, Y., Tan, X., Li, B., Zhang, G., Qin, T., Zhao, S., Shen, Y., Zhang, W.-Q., and Liu, T.-Y. (2021). Adaptive text to speech for spontaneous style. In *Proceedings of Interspeech 2021*, pages 4668−4672. https://doi.org/10.21437/Interspeech.2021-584

Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A.W., and Tokuda K. (2007). The HMM-based speech synthesis system version 2.0*,* In *Proceedings of ISCA SSW6*, pages 294–299. Bonn, Germany.

8. Language Resource References

Yamagishi, J., Veaux, C., and MacDonald, K. (2019). *CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92), [sound].* University of Edinburgh. The Centre for Speech Technology Research (CSTR). https://doi.org/10.7488/ds/2645