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

Institute of the Estonian Language, Tallinn, Estonia

LREC 2022 Marseille Palais du Pharo June 20-25, 2022

Introduction

Synthetic voices are increasingly used in applications that require a conversational speaking style, such as chatbots and automatic dubbing.
RESEARCH QUESTION. Which type of training data yields the most suitable synthetic voices for conversational applications?
HYPOTHESIS. Audiobook dialogues are a suitable source for training conversational style synthetic voices.

Procedure

STEP 1: Creation of three experimental corpora (Estonian, speaker PT, 99,500 characters, 48 kHz, 16 bit, Mono, 70dB):

1. **Character Speech Corpus (CHAR):** dialogues extracted from the PT fiction audiobook corpus [1]
2. **Narrator Speech Corpus (NARR):** speech extracted from the PT fiction audiobook corpus [1] excluding dialogues
3. **Neutral Speech Corpus (NEU):** utterances extracted from the PT neutral sentence-based corpus [2]

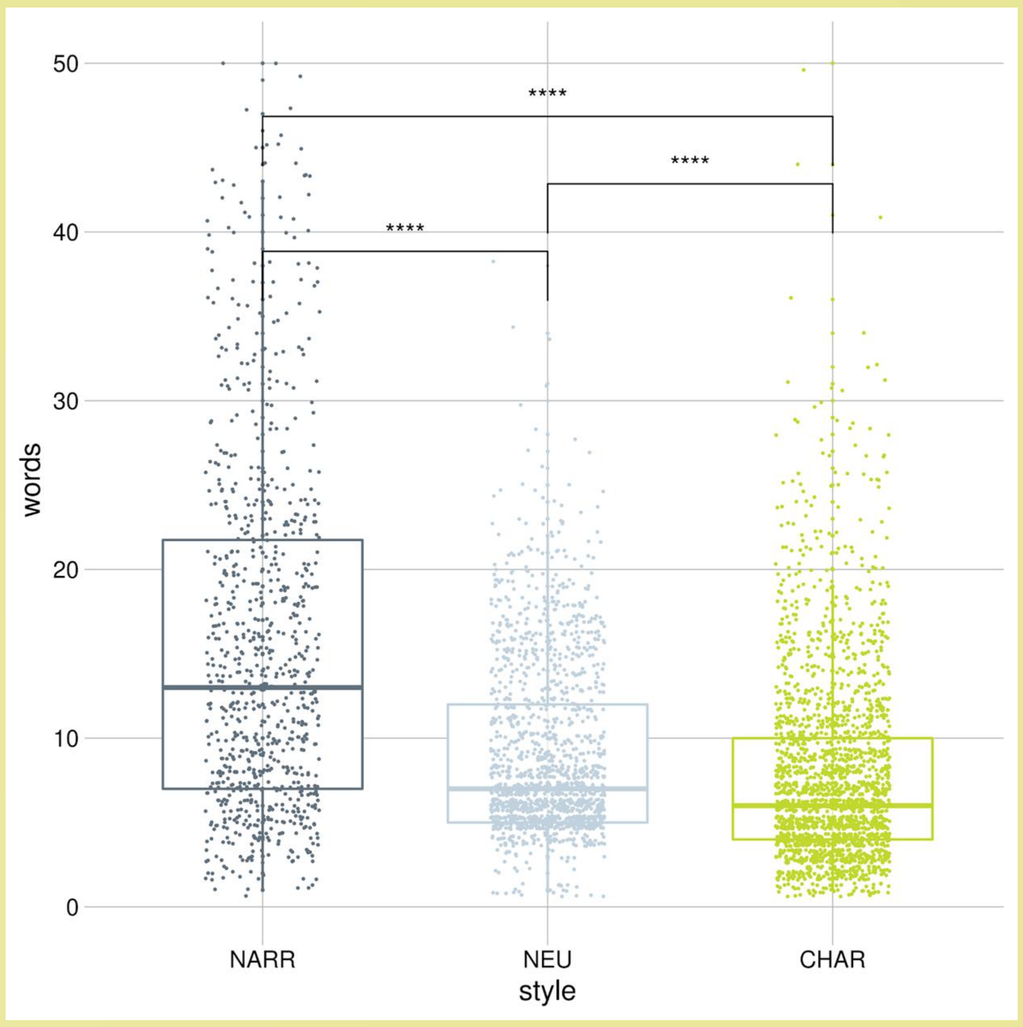

Figure 1. The number of words per sentence in the three corpora.

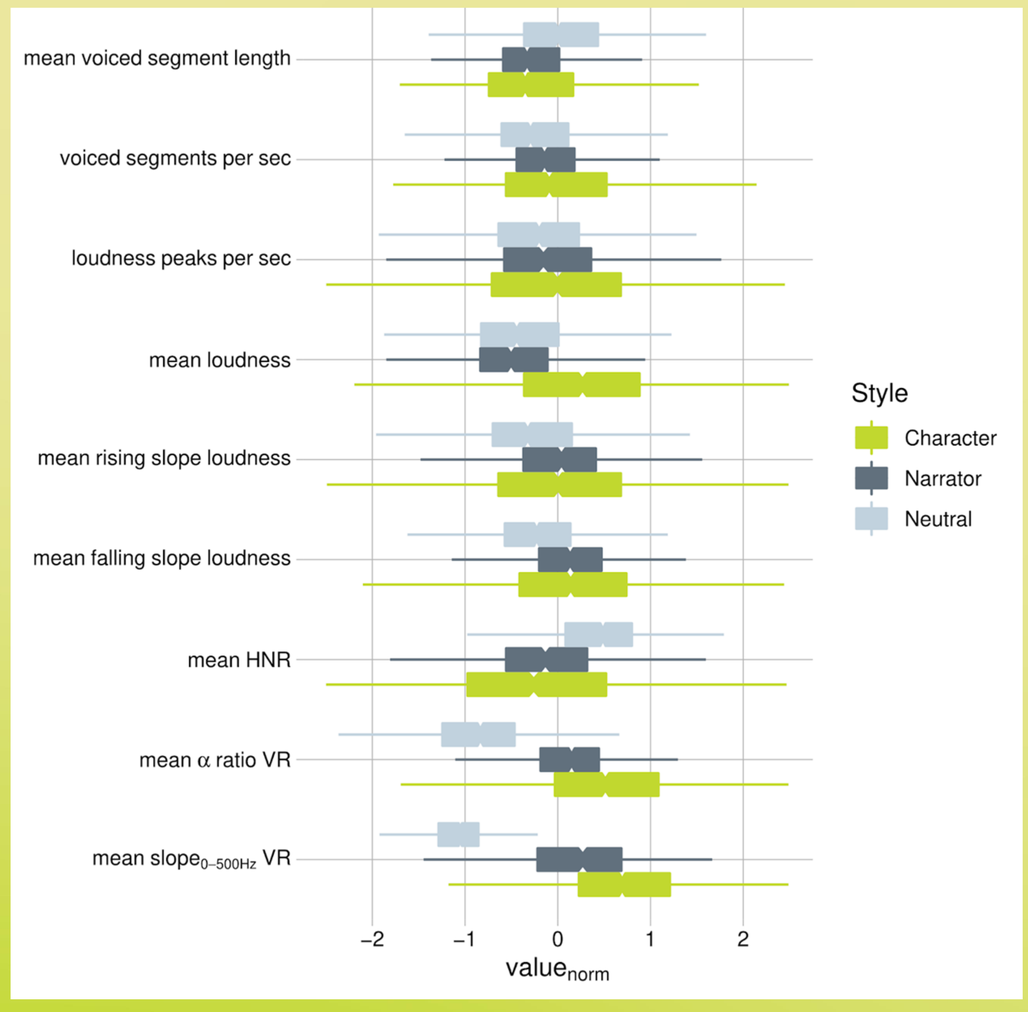

Figure 2. Distinctive acoustic parameters of the three corpora extracted with eGeMAPs [3].
NEU: slower tempo, fewer rapid changes in loudness, a harmonic voice.
CHAR: fastest and loudest speech, rougher and breathier voice.
NARR: intermediate between NEU and CHAR.

STEP 2: Training of nine synthetic voices, using the three corpora (NARR, CHAR, and NEU) and three existing text-to-speech synthesisers:
S1 HMM-based statistical–parametric TTS, phoneme-based approach [4]
S2 HMM-based statistical–parametric TTS, grapheme-based approach [5]
S3 neural network-based TransformerTTS [6]

STEP 3: Synthesising six real customer service chatbot speech turns for evaluation using the nine synthetic voices. For example,
Mul on hea meel, et sain teile abiks olla! [I am glad to have been of assistance!]

STEP 4: Evaluation of the synthesised utterances
A web-based listening test with eight men and eight women (aged 31–65, *M* = 46.0, *SD* =11.1).
The listeners evaluated the suitability of the speaking style on a 7-point Likert scale, where 1 = not suitable at all ... 7 = very suitable.

Results

Voices trained on the NEU 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 NEU and NARR Corpus. Listeners found the voices trained on the CHAR Corpus to be the least suitable, regardless of synthesis technique.

Figure 3. Evaluation results.

Conclusion

The voices trained on the CHAR Corpus received the lowest, and those trained on the NEU 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 CHAR Corpus. The next step will therefore be the creation of a more uniform, balanced, and representative audiobook dialogue corpus.

References
[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] Eyben, F., Scherer, K., Schuller, B., Sundberg, J., Andre, E., Busso, C., ... 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.
[4] https://github.com/ikiissel/synthts_et
[5] https://github.com/CSTR-Edinburgh/Ossian
[6] https://github.com/as-ideas/TransformerTTS

INSTITUTE OF THE ESTONIAN LANGUAGE
European Union European Regional Development Fund
Investing in your future