TURNING PODCASTS INTO A TRAINING CORPUS FOR CONVERSATIONAL TEXT-TO-SPEECH SYNTHESIS

Heete Sahkai, Liisi Piits, Liis Ermus, Indrek Hein, Meelis Mihkla, Indrek Kiissel, Kristjan Suluste, Egert Männisalu, Rene Altrov, Hille Pajupuu, Jaan Pajupuu

Institute of the Estonian Language
heete.sahkai@eki.ee

ABSTRACT

Text-to-speech synthesis (TTS) is increasingly used in applications that require a conversational speaking style, such as voice-enabled chatbots. This gives rise to the need to develop standard and affordable solutions for obtaining conversational training data. This paper describes the creation of a three-hour TTS training corpus for Estonian, using found podcast data, existing automatic speech recognition and transcription-editing software, and a simplified transcription-editing protocol. The corpus was evaluated in comparison with a corpus of read-aloud sentences recorded by the host of the podcast. The evaluation results showed that conversational utterances synthesised with the voices based on the podcast corpus were indeed perceived as representing a more spontaneous speaking style than the utterances synthesised with the voices based on the read-aloud data. Perceived spontaneity was increased by the presence of filled pauses and disfluencies in the stimulus utterances.

Keywords: TTS, spontaneous speech corpus, conversational speaking style, podcasts, Estonian

1. INTRODUCTION

Synthetic voices are increasingly used in applications that require a conversational speaking style, such as voice-enabled chatbots. This gives rise to the need for conversational training data. Obtaining phonetically representative conversational data from a single speaker with sufficient recording quality presents several challenges: what sources can provide such data; how to segment the data into utterances; how to treat the characteristic features of spontaneous speech like filled pauses, backchannels, disfluencies, reduced pronunciation, laughter, etc.

In previous attempts to train conversational synthetic voices, different types of training data have been used: scripted conversational lines or dialogues performed in a studio [1], [2], studio-recorded unscripted dialogues [2]–[4] and monologues [5], [6], existing podcasts [7], [8]. While [3] found that utterances synthesised with voices trained on spontaneous data were perceived as more conversational only when they contained discourse markers and filled pauses, [7] and [8] found that voices based on spontaneous data were generally perceived as more appropriate and spontaneous. Conversational data has also been used in combination with read-aloud data [3] or models based on such data [4], [7]–[9]. This has been found to improve the quality of synthetic speech while retaining its conversational character [3], [7].

The segmentation of unscripted conversational data into utterances has been done manually [3] or using a breath detection method [7], [9].

Regarding the treatment of disfluencies in the training data, the options are to exclude the utterances containing them, or to include them either without annotation or with a more or less detailed annotation [10]. When disfluent utterances are included, a selection must still be operated; for example, [3] retained filled pauses, backchannels, and discourse markers, but excluded utterances with imitations, word fragments, mispronunciations, heavily reduced pronunciations, mumbling and laughter. [7] and [10] found that a voice trained only on fluent utterances was perceived as equally conversational but had better quality than voices trained on disfluent data.

A further question is whether features like filled pauses should be added to input text at the time of synthesis. [3] found that utterances with and without fillers were perceived as equally conversational, while [10] found that filled pauses made synthetic speech sound more authentic.

The goal of the present study is to move towards developing an optimal and affordable workflow for obtaining spontaneous training data. A second goal is to test whether voices based on such data are perceived as having a more spontaneous speaking style than those based on read data, and whether the perception of spontaneity depends on the presence of disfluencies in the input sentences, as the previous evaluation results are mixed in these regards.

We describe the compilation of a spontaneous Estonian TTS corpus. We used found data (podcasts) as an affordable solution, as recording a corpus of conversational data in a studio is costly and time and effort consuming (Section 2). The data was transcribed, edited, annotated, and segmented into utterances using existing automatic speech recognition (ASR) and transcription-editing tools (Section 3). The corpus was evaluated in comparison with a corpus of read-aloud sentences recorded by the host of the podcast (Section 4). We conclude that the resulting corpus does provide suitable training data for conversational style TTS (Section 5).

2. THE DATA

We used a podcast series as a source of conversational speech. The series was chosen from among a set of candidate podcasts and radio shows, based on the following criteria: most of the episodes were recorded by the same speakers, allowing to obtain more material per speaker; there was no background music or noise; the podcast was recorded by a man and a woman, allowing a more reliable speaker recognition, as the ASR system was not able to diarise several male or female speakers in the same recording; there were at most two speakers (ASR makes more diarisation mistakes when there are several speakers). The episodes were obtained through a free public portal for podcasts and radio shows (podcast.ee). We used 11 episodes with a duration of 8.68 hours.

The podcast series is titled “Intimately about private life”[1]. All the episodes are hosted by a married couple, who are not professional radio hosts. The podcast is unscripted and represents an informal spoken register. The topic of the podcast are relationship issues, which are discussed very openly.

An agreement was signed with both presenters, who gave the permission to use their voices for synthesis and for compiling a publicly available corpus.

3. PROCESSING THE MATERIAL

The sound files were transcribed using the Tallinn University of Technology Estonian ASR system[2] [11], [12]. The system performs speaker diarisation, speech-to-text, punctuation restoration to correspond to the written language, text normalisation (reduced forms are corrected). The system omits filled pauses and disfluencies, like interrupted words, from the output text. The system returns the transcription in several formats, from which trs and json were used.

The transcriptions were checked and corrected. The processing started by dividing the transcriptions into shorter chunks in one speaker’s text according to the time codes in the json file, using a dedicated script. Then the transcription in the trs-files was checked manually, using the program Transcriber[3]. Recognition errors were corrected. Punctuation was matched with the structure of the speech (commas marked short pauses in an utterance, periods marked utterance boundaries).

A protocol was established for the annotation of two types of filled pauses (vowel- and consonant-based), backchannels, disfluencies, and laughter. The purpose of the annotation was to gain control over these phenomena (cf. [10]), and to test their influence on the perception of spontaneity.

Utterances containing overlapping speech, laughter, mumbling, and imitations were omitted.

It took about one hour to work through 10 minutes of the transcription.

After the correction and annotation, a second verification and the final selection of utterances was made by another annotator, using an internally developed web-interface. Finally, the trs and sound file were spliced into correspondent one-utterance files (txt and wav). The process resulted in two TTS corpora, one for each speaker. The corpus used for evaluation consisted of 2269 utterances by the female speaker with a duration of 3.69 hours. The corpus will be referred to as the SPON corpus.

4. EVALUATION

The SPON corpus was evaluated in order to answer the following questions: 1) Do listeners perceive a voice based on spontaneous training data as having a more spontaneous speaking style than a voice based on read-aloud data? 2) Does the presence of features that are unique to spontaneous speech affect the perception of spontaneity? 3) Are the evaluation results stable across different synthesis techniques? 4) Is the material sufficient and suitable for different synthesis techniques?

For the purpose of evaluation, a control corpus of read speech was first recorded by the same speaker (4.1). Next, three synthesis methods were used to train a synthetic voice on each corpus (4.2). Then input texts were selected and stimuli were synthesised with the different voices (4.3). The stimuli were then subjected to evaluation (4.4). The evaluation results are presented in Section 4.5.

4.1. Creation of the control corpus

In order to evaluate the SPON corpus against the baseline of read-aloud data a control corpus (READ corpus) was first recorded by the female host of the podcast. The corpus consisted of 1192 phonetically representative read-aloud sentences and had a total duration of 2.06 hours.

4.2. Training of the synthetic voices

Based on each of the two corpora, SPON and READ, three synthetic voices were trained, using different techniques (S1, S2, S3):

S1 uses HTS 2.0, a statistical-parametric TTS technique based on hidden Markov models [13]. The grapheme-based technique uses the language independent text processing libraries of the Ossian-TTS[4] as the front end of the HTS system [14];

S2 uses Merlin, a Neural Network based TTS [15]. The system relies on the Theano numerical computation library. To convert text into full-context labels, an internally developed front-end text processor was used [16];

S3 uses TransformerTTS[5], a TTS solution using a transformer-based neural network model. Spectrograms generated by the model were converted into waveforms using a HiFiGAN vocoder pre-trained on the LJSpeech dataset.

We failed to achieve a satisfying result with S3 on the SPON corpus, thus in total five voices were trained: S1READ, S1SPON, S2READ, S2SPON, and S3READ. S3READ was included in the comparison as a baseline as voices trained with S3 using read-aloud data have received the highest scores in previous evaluations [17].

4.3. The stimuli

Spontaneous speech synthesis should be evaluated using transcriptions of spontaneous speech (see [5]). We used the podcast that served to compile the SPON corpus and selected four 150–200 character passages from the episodes that were not included in the corpus. Each passage was used to create two input texts: 1) an exact transcription of the passage, with filled pauses, repetitions, self-repairs, filler words, etc., to be referred to as ‘natural’ text, and 2) an edited version of the passage, with disfluencies removed, referred to as ‘clean’ text (see Datasets[6]). Each of the eight input texts was synthesised with the five synthetic voices, giving 40 stimuli.

4.4. Testing

The listeners were asked to evaluate each stimulus for how spontaneous it sounds on a 7-point scale, where 1 = not spontaneous at all … 7 = spontaneous. The listeners were 10 women (aged 37–56, *M* = 43.0 years, *SD* = 6.0) and 8 men (aged 34–55, *M* = 43.3 years, *SD* = 6.8). All scores for each listener were normalised using the formula

(1) $y=(x-X)/s$,

where x is the score, X is the mean of the listener’s scores, and s is the standard deviation of the listener’s scores. We classified the performances with scores above zero as spontaneous, and those with scores below zero as not spontaneous.

To find out the degree of agreement among the listeners (inter-rater reliability), the intra-class correlation coefficient (ICC2k) was calculated using the ‘psych’ package in R [18]. A Welch Two Sample *t*-test was used to determine whether the average spontaneity scores differ significantly depending on the training corpus, the synthesis technique, and the type of input text [19].

4.5. Evaluation results

A good to excellent degree of agreement was found between the listeners’ ratings. The average measure ICC2k was 0.90 with a 95% confidence interval from 0.85 to 0.94 $F(39, 663) = 9.8$, $p < 0.0001$).

The results of the *t*-test revealed that the voices trained on the SPON corpus were significantly more spontaneous than the voices trained on the READ corpus: $M_{\text{S1SPON}} = 0.01$ vs. $M_{\text{S1READ}} = -0.76$, $t(267) = -7.46$, $p < 0.0001$; $M_{\text{S2SPON}} = 0.30$ vs. $M_{\text{S2READ}} = -0.13$, $t(286) = -4.18$, $p < 0.0001$ (see Fig. 1).

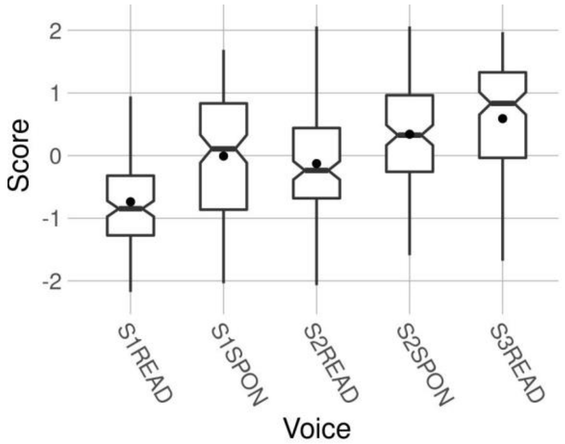

Figure 1: The spontaneity scores by synthetic voices (S1, S2, S3 – synthesis techniques; READ – trained on the read-aloud corpus, SPON – trained on the spontaneous corpus).

Taking into account the synthesis technique, the training corpus, and the type of input text, listeners gave the following spontaneity scores from the lowest to the highest: S1READ_natural -1.01; S2READ_natural -0.50; S1READ_clean -0.50; S1SPON_clean -0.12; S2SPON_clean 0.10; S1SPON_natural 0.14; S2READ_clean 0.24; S3READ_natural 0.51; S2SPON_natural 0.51; S3READ_clean 0.62 (see Fig. 2).

The three voices trained on the READ corpus received higher scores for the clean than for the natural input texts (significant difference for S1 and S2, no significant difference for S3): $M_{\text{S1READ\_clean}} = -0.50$ vs. $M_{\text{S1READ\_natural}} = -1.01$, $t(141) = 4.40$, $p < 0.0001$; $M_{\text{S2READ\_clean}} = 0.24$ vs. $M_{\text{S2READ\_natural}} = -0.50$,

$t(138) = 5.59$, $p < 0.0001$; $M_{\text{S3READ\_clean}} = 0.63$ vs. $M_{\text{S3READ\_natural}} = 0.51$, $t(142) = 0.73$, $p = 0.465$.

The two voices trained on the SPON corpus received higher scores for the natural input texts (significant difference in the case of S2): $M_{\text{S1SPON\_clean}} = -0.12$ vs. $M_{\text{S1SPON\_natural}} = 0.14$, $t(140) = -1.63$, $p = 0.106$; $M_{\text{S2SPON\_clean}} = 0.10$ vs. $M_{\text{S2SPON\_natural}} = 0.51$, $t(137) = -2.94$, $p = 0.004$.

The scores of S1SPON and S2SPON did not differ significantly in case of the clean input text: $M_{\text{S1SPON\_clean}} = -0.12$ vs. $M_{\text{S2SPON\_clean}} = 0.10$, $t(132) = -1.45$, $p = 0.151$. By contrast, S2SPON got a significantly higher spontaneity score for natural texts: $M_{\text{S1SPON\_natural}} = 0.14$ vs. $M_{\text{S2SPON\_natural}} = 0.51$, $t(142) = -2.42$, $p = 0.016$.

The performances of both natural and clean texts by S3READ received equal scores to the natural texts performed by S2SPON: $M_{\text{S2SPON\_natural}} = 0.51$ vs. $M_{\text{S3READ\_natural}} = 0.51$, $t(142) = 0.01$, $p = 0.995$; $M_{\text{S2SPON\_natural}} = 0.51$ vs. $M_{\text{S3READ\_clean}} = 0.63$, $t(142) = -0.71$, $p = 0.473$.

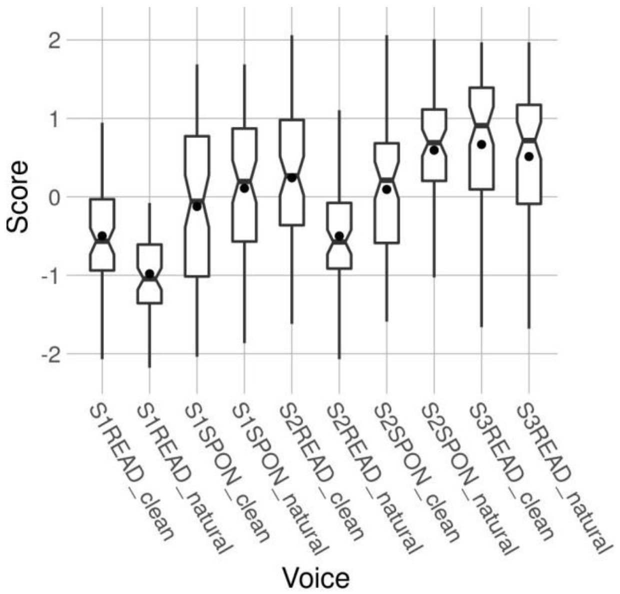

Figure 2: Spontaneity scores by the synthetic voices and the type of input text (S1, S2, S3 – synthesis techniques; READ – trained on the read-aloud corpus, SPON – trained on the spontaneous corpus; natural – input texts with fillers and disfluencies, clean – input texts without fillers and disfluencies).

5. DISCUSSION AND CONCLUSION

Both voices trained on the SPON corpus were perceived as significantly more spontaneous than the voices trained on the READ corpus using the same techniques (Fig. 1). The presence of fillers and disfluencies in the input text increased the perceived spontaneity of the voices trained on the SPON corpus, confirming the results of [10]. For the voices trained on the READ corpus, the effect was reversed: fillers and disfluencies in the input text decreased the perceived spontaneity (Fig. 2). Our results thus support the use of spontaneous training data as well as the insertion of fillers and disfluencies into synthetic speech by one of the methods in [10].

As to the performance of the different synthesis techniques, the SPON corpus was suitable for creating synthetic voices with two of the techniques, S1 and S2, while S3 did not yield results. Due to the characteristics of spontaneous speech (reduced articulation, fillers, disfluencies), this type of synthesis technique might require a larger training corpus (cf. [1]), or a combination of spontaneous and read data (see [3], [7]). From the two successful techniques, S2 received significantly higher scores for the natural input texts than S1. Equal scores were received by the baseline voice trained with S3 on the READ corpus (Fig. 2).

Regarding the methods for obtaining spontaneous training data, the results suggest that a relatively small amount of spontaneous speech (three hours) allows to train synthetic voices with certain methods. Previous attempts have been based on larger corpora, cf. [1], [7], [8].

The most time-consuming task was the manual correction and annotation of the data. This could be alleviated by developing an ASR tool for an accurate transcription of spontaneous speech, including filled pauses, disfluencies, etc. Another possible future avenue is the improvement of the annotation protocol based on the analysis of the synthesis results.

6. ACKNOWLEDGEMENTS

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 the basic governmental financing of the Institute of the Estonian Language from the Estonian Ministry of Education and Research (project EKI-BAAS-2023/2 Speech studies at the service of speech technology). We wish to thank our annotators and informants.

7. REFERENCES

[1] Zandie, R., Mahoor, M. H., Madsen, J., Emamian, E. S. 2021. RyanSpeech: A Corpus for Conversational Text-to-Speech Synthesis. *Interspeech 2021*, 2751–2755. https://doi.org/10.21437/Interspeech.2021-341

[2] Adigwe, A. O., Klabbers, E. 2022. Strategies for developing a Conversational Speech Dataset for Text-To-Speech Synthesis. *Interspeech 2022*, 2318–2322. https://doi.org/10.21437/Interspeech.2022-10802

[3] Andersson, S., Yamagishi, J., Clark, R. A. J. 2010. Utilising spontaneous conversational speech in HMM-based speech synthesis. *Proc. 7th ISCA Workshop on Speech Synthesis (SSW 7)*, 173–178.

[4] Huang, Y.-C., Wu, C.-H., Chen, Y.-Y., Shie, M.-G., Wang, J.-F. 2017. Personalized Spontaneous Speech Synthesis Using a Small-Sized Unsegmented Semispontaneous Speech. *IEEE/ACM Trans. Audio Speech Lang. Process*, *25*(5), 1048–1060. https://doi.org/10.1109/TASLP.2017.2679603

[5] Székely, E., Edlund, J., Gustafson, J. 2020. Augmented Prompt Selection for Evaluation of Spontaneous Speech Synthesis. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 6368–6374. https://aclanthology.org/2020.lrec-1.782

[6] Lameris, H., Mehta, S., Henter, G. E., Gustafson, J., Székely, É. 2022. Prosody-controllable spontaneous TTS with neural HMMs (arXiv:2211.13533). arXiv. http://arxiv.org/abs/2211.13533

[7] Székely, É., Henter, G. E., Beskow, J., Gustafson, J. 2019. Spontaneous Conversational Speech Synthesis from Found Data. *Interspeech 2019*, 4435–4439. https://doi.org/10.21437/Interspeech.2019-2836

[8] Yan, Y., Tan, X., Li, B., Zhang, G., Qin, T., Zhao, S., Shen, Y., Zhang, W.-Q., Liu, T.-Y. 2021. Adaptive Text to Speech for Spontaneous Style. *Interspeech 2021*, 4668–4672. https://doi.org/10.21437/Interspeech.2021-584

[9] Székely, É., Henter, G. E., Gustafson, J. 2019. Casting to Corpus: Segmenting and Selecting Spontaneous Dialogue for Tts with a Cnn-lstm Speaker-dependent Breath Detector. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6925–6929. https://doi.org/10.1109/ICASSP.2019.8683846

[10] Székely, É., Eje Henter, G., Beskow, J., Gustafson, J. 2019. How to train your fillers: Uh and um in spontaneous speech synthesis. *10th ISCA Workshop on Speech Synthesis (SSW 10)*, 245–250. https://doi.org/10.21437/SSW.2019-44

[11] Alumäe, T., Tilk, O., Asadullah. 2018. Advanced Rich Transcription System for Estonian Speech. *Human Language Technologies – The Baltic Perspective*, 1–8. https://doi.org/10.3233/978-1-61499-912-6-1

[12] Olev, A., Alumäe, T. 2022. Estonian Speech Recognition and Transcription Editing Service. *Baltic Journal of Modern Computing*, *10*(3). https://doi.org/10.22364/bjmc.2022.10.3.14

[13] Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. W., Tokuda, K. 2007. The HMM-based Speech Synthesis System (HTS) Version 2.0. *SSW6-2007*, 294–299

[14] Vainio, M., Grönroos, S.-A., Smit, P., Suni, A., Watts, O. 2014. *Deliverable D2.2. Description of the final version of the new front-end*. https://simple4all.org/wp-content/uploads/2014/11/Simple4All_deliverable_D2.2.pdf

[15] Wu, Z., Watts, O., King, S. 2016. Merlin: An Open Source Neural Network Speech Synthesis System. *9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, 202–207. https://doi.org/10.21437/SSW.2016-33

[16] Kiissel, I. 2022. *Merlinil põhinev eesti keele kõnesüntesaator [Merlin based Estonian speech synthesizer]*. https://github.com/ikiissel/mrln_et

[17] Piits, L., Pajupuu, H., Sahkai, H., Altrov, R., Ermus, L., Tamuri, K., Hein, I., Mihkla, M., Kiissel, I., Männisalu, E., Suluste, K., Pajupuu, J. 2022. Audiobook Dialogues as Training Data for Conversational Style Synthetic Voices. *Proceedings of the Language Resources and Evaluation Conference*, 1047–1053. https://aclanthology.org/2022.lrec-1.112

[18] Revelle, W. 2022. *psych: Procedures for Psychological, Psychometric, and Personality Research* (R package version 2.2.9). Northwestern University. https://CRAN.R-project.org/package=psych

[19] R Core team. 2022. *R: A Language and Environment for Statistical Computing* (4.2.2). R Foundation for Statistical Computing. https://www.R-project.org/

[1] https://podcast.ee/show/intiimselt-eraelust/

[2] http://bark.phon.ioc.ee/webtrans/

[3] https://sourceforge.net/projects/trans/files/transcriber/

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

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

[6] The datasets generated and analysed during the study: https://figshare.com/projects/Turning_podcasts_into_a_training_corpus_for_conversational_text-to-speech_synthesis/156248