Interspeech 2024
1-5 September 2024, Kos, Greece

Enabling Conversational Speech Synthesis using Noisy Spontaneous Data

Liisa Rätsep[1], *Rasmus Lellep*[1], *Mark Fishel*[1]

[1]University of Tartu, Estonia

liisa.ratsep@ut.ee, rasmus.lellep@ut.ee, mark.fisel@ut.ee

Abstract

In recent years, the quality of text-to-speech models has increased significantly, but most text-to-speech solutions are trained on datasets of read speech and do not cover conversational speaking styles due to the lack of suitable training data. This paper explores options for creating multi-style speech synthesis using speech recognition datasets that contain samples of spontaneous speech and dialogues but may also include background noise and an insufficient number of samples per speaker. We develop an Estonian multi-speaker TTS system that increases prosodic variability on conversational inputs while still being able to synthesize read speech. We show that our proposed approach can be used to train models that can be controlled to produce conversational speech with little compromise on audio quality. We also highlight a potential multilingual use case to achieve cross-lingual speaker and style transfer to low-resource languages that lack stylistically diverse speech corpora.

Index Terms: conversational speech synthesis, style transfer, low-resource settings

1. Introduction

Thanks to the emergence of various neural text-to-speech (TTS) architectures in recent years, the quality of synthesized speech has increased significantly. However, most publicly available TTS systems are trained on datasets of read speech and cannot model the prosody of unseen speaking styles, such as conversational speech [1]. While there are commercial TTS voices that can produce conversational speech, they are often trained on proprietary datasets and support a very limited set of languages.

Many languages, however, have speech datasets that contain some spontaneous speech for developing automatic speech recognition (ASR) systems. While this type of found data could be a useful low-cost alternative for modeling conversational speech, ASR datasets also contain noise and limited samples per speaker, which make them suboptimal for training TTS models. Additionally, using these speakers as target speakers may be considered unethical as they have often not agreed to be imitated by text-to-speech. While there have been previous works regarding conversational TTS that use the spontaneous speech samples in found data corpora [1, 2, 3], they often focus on English, use carefully selected high-quality recordings, such as podcasts, and train a model to be used exclusively in conversational settings.

This work explores options for using existing speech recognition datasets that contain samples of spontaneous speech and dialogues but also include background noise. We propose several changes to a standard non-autoregressive multi-speaker Transformer-based TTS architecture to increase prosodic variability and expressiveness in conversational settings via explicit style conditioning. We evaluate our approach with existing Estonian datasets and show that our proposed method can be used to train multi-style models that produce both conversational and read speech and transfer these styles between different speakers with little compromise on overall speech quality.

Additionally, we explore a multilingual use case of our approach by adding read speech data for Võro – a related low-resource language. We demonstrate that our method could be potentially used to achieve cross-lingual speaker and style transfer to low-resource languages, which may lack any stylistically diverse speech corpora.

2. Related work

As conversational text-to-speech corpora are scarce, existing research has proposed several approaches to mitigating this problem using found data.

[1, 2, 3] have suggested using podcasts with automatic transcriptions. [2] propose a modified Transformer-based architecture with filled pause prediction initially trained on read speech and fine-tuned on conversational data. They propose training the filled pause prediction, variation adaption, and timbre adaptation separately in different stages and on potentially different datasets. [4] propose an approach based on an adapted version of FastSpeech 2 [5], which is pre-trained on large amounts of low-quality spontaneous data in Mandarin and later fine-tuned high-quality conversational speech. While these contributions result in a TTS model for conversational speech, our approach results in a multi-style model that covers both read speech as well as conversational speech style.

[3] constructed a dataset of questions-answer pairs and added it to read data when training a multi-speaker FastPitch [6] model. While they noticed some improvement in synthesizing conversational speech, the model did not learn to transfer the style to other speakers effectively.

Conversational neural text-to-speech for Estonian or low-resource languages is still relatively unexplored. [7] trained several TTS models on dialogue portions of Estonian audiobooks, but there have been no models trained on spontaneous conversational data.

3. Method

Our core approach consists of modifying the multi-speaker version of FastPitch [6], a non-autoregressive Transformer-based encoder-decoder architecture with convolutional variance predictors for character durations and pitch contours. The original multi-speaker model is trained with global speaker embeddings, which are added to encoder input tokens. Our approach is to

10.21437/Interspeech.2024-2403

prepend a speaker embedding only to the decoder input, and to train separate style embeddings, adding them to duration and pitch predictor inputs.

Additionally, we experiment with modifications to backpropagation where the gradient calculation from pitch and duration loss functions is restricted to variance predictor weights and style embeddings but not the encoder weights. This modification was designed to ensure that the model learns to pay attention to the style embedding and not predict prosody only based on the input text. It also allows us to train the encoder and decoder separately from variance predictors using potentially different datasets to avoid mel-spectrogram loss being optimized on low-quality data. Moreover, this approach may be useful to ensure that data from certain speakers can still be used for modeling prosody without risking their timbre being imitated or leaking into the embeddings of other speakers.

During our initial experiments, we evaluated other variations of speaker and style parametrization. To explore whether parametrization is required on the encoder level, we experimented with using speaker embeddings as inputs to variance predictors and the decoder. We found that there was no noticeable negative effect on target speaker similarity when compared to the original multi-speaker FastPitch.

We also explored using global style embeddings as additional input for the encoder instead of only the variance predictors. Both models generated different prosody depending on the style embedding value with no noticeable difference in quality. Therefore, as global style embeddings would cause the speaker and style parameters to become entangled, we opted not to use any embeddings on the encoder level.

4. Experiments

4.1. Data

For comparability, our dataset selection for read speech followed our previous work on Estonian TTS [8]. We used the Speech Corpus of Estonian News Sentences [9]. The dataset contains recordings of news texts read by four university students, which we consider neutral in style. Additionally, we used audiobook corpora [10, 11] read by professional actors, representing the style of fiction.

For spontaneous data, we used the ERR2020 corpus [12], which contains manually transcribed Estonian Public Broadcast recordings and is designed for training ASR systems for closed captioning. We used the radio interviews subset of the corpus, which contains 100 hours of speech from 827 speakers. The audio quality is generally very good, but there are also sections of phone interviews, samples where people talk over each other, and many speakers with very few samples.

For low-resource experiments, we added read speech data from two Võro speakers[1]. The datasets contained two speakers reading an identical set of sentences in a neutral style.

The audio was sampled at 22050 Hz, and we used a frame size of 1024 and a hop length of 256 for mel-spectrogram conversion. The audio and text alignment for both languages was generated using the Estonian alignment model [13] as our previous work has demonstrated that it can produce a high alignment quality also for Võro [14]. Excessive pauses were trimmed based on the alignment information.

A random subset of 100 sentences per speaker of read speech datasets and 100 sentences from the entire ERR2020 corpus were excluded for evaluation purposes. Pitch contours were normalized using speaker-specific mean and standard deviation values to make predicted pitch values interchangeable between speakers. 87 speakers with less than 10 training samples were excluded to ensure the mean and standard deviation values were reliable. We removed sentences that had durations outside the range of 0.3 to 17.5 seconds and sentences where the transcription included non-normalized numbers, abbreviations, or symbols. Final training data sizes after filtering can be seen in Table 1.

Table 1: *Training dataset sizes after filtering.*

Style	Speakers	Samples	Duration (h)
Neutral (et)	4	22488	35.5
Fiction (et)	6	45538	58.6
Conversational (et)	740	32663	58.1
Neutral (vro)	2	2061	3

4.2. Models and Training

4.2.1. Monolingual Models

As a baseline, we trained a standard multi-speaker FastPitch model with global speaker embeddings using only read speech data ("multi-speaker") and a version with added spontaneous conversational data ("multi-speaker + spon.") similar to [3]. Additionally, we trained a model without style embeddings where the speaker embedding was prepended only to decoder inputs (multi-speaker decoder) using all data. The goal was to evaluate the impact of using style conditioning as opposed to letting the model automatically predict prosodic features based on just input text.

To evaluate our method, we trained a model with explicit style parametrization ("multi-style") and a model with restricted backpropagation ("multi-style, restricted") using both read and spontaneous data. Finally, we trained a model where the encoder and decoder were trained only on read data, and the variance predictors were trained separately on all data ("multi-style, 2-stage").

4.2.2. Multilingual Models

For the multilingual use case with the low-resource language Võro, we trained a model with style parametrization and no backpropagation restrictions using all datasets (the "multi-style" approach). This model was trained without providing any target language information to the model. During initial experiments, we also explored using global language embeddings as additional encoder inputs but found that it made the speaker and style embeddings too dependent on the target language, which had a negative effect on cross-lingual transfer.

4.2.3. Training Configuration

For training, we used a modified open-source implementation of FastPitch[2]. All models were trained for 150k steps (approximately 2.5 days). Each model was trained on a single NVIDIA A100 40GB GPU using a batch size of 128000 frames. The

[1]Võro datasets Sulev and Hellä: `https://koneveeb.ee/korpused/`

[2]Code and hyperparameter configuration: `https://github.com/TartuNLP/TransformerTTS/tree/interspeech2024`

Table 2: *Evaluation results with 95% confidence intervals. The "read speech" scores refer to samples synthesized with the original target speaker, whereas "read speech (style transfer)" refers to using target speakers from a different read speech style. Highest scores are highlighted.*

	Read speech		Read speech (style transfer)		Conversational	
Method	MOS	StyleMOS	MOS	StyleMOS	MOS	StyleMOS
GT	4.5 ± 0.07	4.39 ± 0.09	-	-	4.66 ± 0.1	4.9 ± 0.06
GT mel + vocoder	4.2 ± 0.09	4.24 ± 0.1	-	-	4.4 ± 0.13	4.85 ± 0.07
Multi-speaker	3.66 ± 0.09	**4.06 ± 0.1**	3.52 ± 0.1	3.38 ± 0.13	3.13 ± 0.1	3.06 ± 0.12
Multi-speaker + spon.	3.53 ± 0.1	3.86 ± 0.11	3.44 ± 0.1	3.35 ± 0.12	3.27 ± 0.11	3.29 ± 0.12
Multi-speaker decoder	3.6 ± 0.1	3.87 ± 0.11	3.53 ± 0.1	3.58 ± 0.12	3.34 ± 0.11	3.58 ± 0.12
Multi-style	3.58 ± 0.09	3.92 ± 0.1	3.5 ± 0.1	3.61 ± 0.11	**3.36 ± 0.1**	**3.74 ± 0.11**
Multi-style, restricted	3.6 ± 0.1	3.92 ± 0.11	3.46 ± 0.1	3.59 ± 0.11	3.31 ± 0.11	3.34 ± 0.12
Multi-style, 2-stage	**3.67 ± 0.1**	3.92 ± 0.11	**3.57 ± 0.1**	**3.68 ± 0.11**	3.35 ± 0.11	3.43 ± 0.12

models contain 96M trainable parameters, and identical hyperparameters were used across all models with only minor differences in the embedding layers and number of embeddings as dictated by the variations in model architecture or the inclusion of conversational and low-resource data.

4.3. Evaluation

For Estonian models, we used a subset of 40 random sentences for each style from the held-out datasets to generate our evaluation samples. The waveforms were created using existing HiFiGAN models[3] [15] trained on VCTK [16] or a ground truth aligned Tacotron 2 [17] samples of LJSpeech [18]. Model selection for each target speaker followed our previous works on the same datasets [8, 14] and was consistent across all models. Additionally, we included original ground truth samples as well as samples reconstructed from ground truth mel-spectrograms. A random speaker ID was selected for synthesizing sentences with target speakers with different corpus styles (style transfer). The same set of sentences and speaker IDs were used for all methods to ensure the comparability of scores.

To assess model quality, we conducted a mean opinion score (MOS) [19] evaluation among native speakers of Estonian. The surveys were conducted in the PCIbex Farm[4] environment. Evaluation samples were split into 24 surveys of 115 samples, and each survey was completed by 5 people. The participants were asked to provide a general score, considering all aspects of speech, and a style-specific score (StyleMOS). For style-specific evaluation, they were asked to imagine themselves in one of three scenarios - listening to a synthesized news article on a website, listening to an audiobook, or listening to a radio show. The participants were instructed only to consider the appropriateness of style and prosody and ignore the pleasantness of speaker timber, audio quality, and TTS artifacts.

5. Results

5.1. Human Evaluation

The evaluation[5] results can be seen in Table 2. Expectedly, when synthesizing conversational texts, we saw the overall speech quality and style improve across all methods when spontaneous speech data was added. While the improvement in general MOS scores was very similar across most methods, the model with style parametrization and unrestricted backpropagation ("multi-style") offers the biggest stylistic improvement.

Simply adding spontaneous data to the multi-speaker baseline provided the smallest benefit. This supports the hypothesis in [3] that existing target speakers in a multi-speaker model will not benefit from additional data as the model will learn separate prosodic features for each speaker. Additionally, the degradation of all scores on read speech styles illustrates that this model offers no benefits in multi-style synthesis.

The evaluation of read speech styles using original target speakers shows a minor decrease in general quality across all methods, with the exception of the model trained in multiple stages that does not use noisy data for training mel-spectrogram prediction ("multi-style, 2-stage"). While the improvement this model provides is not significant, it prevents degradation in synthesized audio quality as we expected. The consistently higher MOS scores of this model compared to the similarly structured restricted variation ("multi-style, restricted") also support this finding. Considering that our ASR dataset is relatively high-quality, this finding may be even more relevant to datasets with more noise.

Style-specific scores (StyleMOS) on original target speakers show a minor decrease compared to the baselines for all methods. This is partially expected, especially for the fiction style where each actor likely had a very expressive distinct style, which, due to style-specific prosody conditioning, averages out between target speakers. However, when synthesizing the same sentences using target speakers from a different read speech domain ("Read speech/style transfer"), we see the best performance from the model trained in multiple stages. In style transfer, all models except for the multi-speaker baseline with additional data offer significant stylistic improvements.

Finally, it should be noted that although the participants were instructed to ignore audio quality and artifacts when evaluating style, the consistently lower ratings of reconstructed ground truth spectrograms compared to ground truth waveforms indicate a bias among our participants to have a lower perception of prosody whenever the audio quality is lower.

5.2. Effects of stylistic parametrization

In case of sentences from the read speech corpora, we saw very similar results on style-conditioned synthesis ("multi-style")

[3]Pretrained HiFiGAN models: `https://github.com/jik876/hifi-gan`

[4]PCIbex Farm: `https://farm.pcibex.net/`

[5]Evaluation samples: `https://tartunlp.github.io/TransformerTTS/interspeech2024/`

and the multi-speaker decoder model. To illustrate whether the models learn to use style embeddings and produce stylistically appropriate prosody, we measured the difference in stylistic evaluations of style transfer when using the target speaker's original style. The results of these evaluations can be found in Table 3.

Table 3: *Relative effect on StyleMOS evaluations when using an incorrect target style ID.*

Method	Read	Conversational
Multi-style	-0.35	-0.25
Multi-style, restricted	-0.26	-0.20
Multi-style, 2-stage	-0.35	-0.24

In all cases, conditioning on the incorrect style resulted in a decreased StyleMOS score. This suggests that the style-conditioned models clearly differentiate between the three styles and actually take advantage of style information without relying only on encoded text for prosody modeling. The multi-speaker decoder models, however, are forced to model this information based on text to cope with not having access to explicit style information. Additionally, we do not observe the model becoming more reliant on style conditioning with restricted backpropagation.

5.3. Qualitative analysis: Võro

To evaluate our Võro model, we interviewed three speakers of Võro who provided their impressions on synthesis quality. While the quality of Võro speech was considered lower than the models published in [14], which were trained for longer using multilingual transfer learning and data augmentation techniques not explored in this work, the participants confirmed that the models produced understandable Võro for both target speakers. However, they noted that the speech was faster, the melody of synthesized sentences was more monotonous and sounded less natural to Võro, whereas the pronunciation was correct. The participants claimed that the Estonian target speakers produced understandable Võro speech with similar quality, whereas generating Võro speech with the equivalent Estonian model did not.

Our testing also showed that the style parametrization generated audibly different prosody in many cases. This suggests that our method may be useful for diversifying TTS for related low-resource languages that lack stylistically diverse speech data. However, the stylistic appropriateness should still be evaluated further in the future with style-specific evaluation sentences. Reintroducing language embeddings may also be necessary to model language-specific prosodic features even for similar languages.

Finally, we synthesized Estonian sentences with all text domains and style parameters with both Võro speakers. We observed no foreign accent or other issues with Estonian pronunciation while producing the correct target speaker timbre. The quality was similar to that of Estonian target speakers. We also observed similar behavior when using style conditioning compared to the monolingual Estonian models.

6. Limitations and Future Work

Our work has several limitations, which we describe below and plan to explore further in future works.

Firstly, our work uses a relatively high-quality dataset of spontaneous speech to evaluate our proposed approach. In the future, we plan to verify and develop it further using additional lower-quality datasets that contain even more noise. Additional Estonian TTS corpora have also recently become available and will be used in future works.

Secondly, we plan to explore more methods of using a limited higher-quality subset of data to optimize mel-spectrogram loss. Alternatively, the same effect can be achieved via custom lower mel loss weights for noisier training samples. Additionally, we will explore options of pre-training or fine-tuning specific parts of the model as alternatives.

Finally, the evaluation of our low-resource language is only a preliminary study. More experimentation is needed to conclude that the approach is reliable in multilingual and low-resource settings. We plan to expand our style-conditioned approach to other related low-resource languages and evaluate it on high-resource languages that could provide additional stylistic and target speaker variety.

7. Conclusions

In this work, we have demonstrated that modifying a non-autoregressive Transformer-based model with style-informed variance predictors and a speaker-informed decoder is an effective way to create a model with speaker-independent style conditioning. By combining lower-quality Estonian spontaneous conversational speech with existing TTS corpora, we achieved a significant stylistic improvement when synthesizing conversational texts while maintaining the ability to synthesize read-style speech. We also demonstrated that our method could benefit multilingual settings, especially for low-resource languages.

8. Acknowledgements

This work was supported by the Estonian Research Council grant PRG2006 (Language Technology for Low-Resource Finno-Ugric Languages and Dialects), the National Programme of Estonian Language Technology grant EKTB92 (Expressive Text-to-Speech Synthesis for New Languages and Users) and the Estonian Centre of Excellence in Artificial Intelligence (EXAI), funded by the Estonian Ministry of Education and Research.

9. References

[1] Éva Székely, G. E. Henter, J. Beskow, and J. Gustafson, "Spontaneous Conversational Speech Synthesis from Found Data," in *Proc. Interspeech 2019*, 2019, pp. 4435–4439.

[2] Y. Yan, X. Tan, B. Li, G. Zhang, T. Qin, S. Zhao, Y. Shen, W.-Q. Zhang, and T.-Y. Liu, "Adaptive Text to Speech for Spontaneous Style," in *Proc. Interspeech 2021*, 2021, pp. 4668–4672.

[3] J. O'Mahony, C. Lai, and S. King, "Combining conversational speech with read speech to improve prosody in Text-to-Speech synthesis," in *Proc. Interspeech 2022*, 2022, pp. 3388–3392.

[4] W. Li, S. Lei, Q. Huang, Y. Zhou, Z. Wu, S. Kang, and H. Meng, "Towards Spontaneous Style Modeling with Semi-supervised Pre-training for Conversational Text-to-Speech Synthesis," in *Proc. INTERSPEECH 2023*, 2023, pp. 3377–3381.

[5] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and high-quality end-to-end text to speech," in *International Conference on Learning Representations*, 2021.

[6] A. Łańcucki, "FastPitch: Parallel text-to-speech with pitch prediction," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6588–6592.

[7] L. Piits, H. Pajupuu, H. Sahkai, R. Altrov, L. Ermus, K. Tamuri, I. Hein, M. Mihkla, I. Kiissel, E. Männisalu, K. Suluste, and J. Pajupuu, "Audiobook dialogues as training data for conversational style synthetic voices," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, Eds. Marseille, France: European Language Resources Association, Jun. 2022, pp. 1047–1053. [Online]. Available: https://aclanthology.org/2022.lrec-1.112

[8] L. Rätsep, R. Lellep, and M. Fishel, "Estonian text-to-speech synthesis with non-autoregressive transformers," *Baltic Journal of Modern Computing*, vol. 10, 01 2022.

[9] M. Fishel, A. Laumets-Tättar, and L. Rätsep, "Speech corpus of Estonian news sentences," https://doi.org/10.15155/9-00-0000-0000-0000-001ABL, 2020.

[10] L. Piits, "Estonian male voice audiobook corpus for speech synthesis," https://doi.org/10.15155/3-00-0000-0000-0000-08BF4L, 2022.

[11] ——, "Estonian female voice audiobook corpus for speech synthesis," https://doi.org/10.15155/3-00-0000-0000-0000-090D4L, 2022.

[12] T. Alumäe, J. Kalda, K. Bode, and M. Kaitsa, "Automatic closed captioning for Estonian live broadcasts," in *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*. Tórshavn, Faroe Islands: University of Tartu Library, May 2023, pp. 492–499. [Online]. Available: https://aclanthology.org/2023.nodalida-1.49

[13] T. Alumäe, O. Tilk, and Asadullah, "Advanced rich transcription system for Estonian speech," in *Human Language Technologies - the Baltic Perspective: Proceedings of the Eighth International Conference*. IOS Press, 2018, pp. 1–8.

[14] L. Rätsep and M. Fishel, "Neural text-to-speech synthesis for võro," in *The 24rd Nordic Conference on Computational Linguistics*, 2023. [Online]. Available: https://openreview.net/forum?id=V5PGSHHJEw

[15] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2020, pp. 17 022–17 033.

[16] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," https://datashare.ed.ac.uk/handle/10283/3443, 2019.

[17] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779–4783.

[18] K. Ito and L. Johnson, "The LJ Speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[19] M. Chu and H. Peng, "An objective measure for estimating MOS of synthesized speech," in *EUROSPEECH 2001, 7th European Conference on Speech Communication*. ISCA, 2001, pp. 2087–2090.