Baltic J. Modern Computing, Vol. 10 (2022), No. 3, pp. 447–456
https://doi.org/10.22364/bjmc.2022.10.3.17

Estonian Text-to-Speech Synthesis with Non-autoregressive Transformers

Liisa RÄTSEP, Rasmus LELLEP, Mark FISHEL

Institute of Computer Science, University of Tartu, Estonia

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

Abstract. While text-to-speech synthesis with non-autoregressive Transformers has achieved state-of-the-art quality for many languages, the methodology of Estonian text-to-speech synthesis has not been revised for neural methods. This paper evaluates the quality of Estonian text-to-speech with Transformer-based models using different language-specific data processing steps. Additionally, we conduct a human evaluation to show how well these models can learn the patterns of Estonian pronunciation, given different amounts of training data and varying degrees of phonetic information. Our error analysis shows that using a simple multi-speaker approach can significantly decrease the number of pronunciation errors, while some information can also be helpful to a smaller extent.

Keywords: speech technology, text-to-speech synthesis, Estonian

1 Introduction

In recent years, non-autoregressive Transformer-based models have achieved high quality in neural text-to-speech (TTS) synthesis while maintaining reasonable inference speeds (Ren et al., 2019, 2021; Łańcucki, 2021). However, most neural TTS methods have been developed with English datasets and rely heavily on language-specific tools such as alignment models and grapheme-to-phoneme converters.

At the same time, existing Estonian TTS research has mainly focused on other methods, such as concatenative synthesis or statistical parametric synthesis (Mihkla et al., 2008; Nurk, 2012). Although there are some Transformer-based Estonian models publicly available, their quality has not been evaluated, and there is a general lack of insight when it comes to best practices with neural methods.

While Estonian orthography is not entirely phonetic, it is highly phonetically motivated. Therefore, it is worth investigating whether neural networks need additional information from grapheme-to-phoneme conversion tools to reduce the phonetic ambiguity in texts or if they can surpass the level of rule-based annotations with sufficient

training data. Additionally, as non-autoregressive models are conditioned on character duration, it should be evaluated whether we can reuse existing Estonian grapheme-to-waveform alignment models in TTS. An alternative option is to extract the duration information from an autoregressive teacher model (Ren et al., 2019), but this approach significantly increases the time and cost of developing new models.

This paper will analyze the performance of state-of-the-art neural TTS methods when applied to Estonian and evaluate the impact of integrating existing Estonian NLP tools into TTS workflows. Our primary goal is to provide a guide to best practices for Estonian neural TTS and highlight the weaknesses that should be tackled in the future. We approach this task by training models on character durations extracted from a teacher model and comparing the speech quality to models using the external aligner by Alumäe et al. (2018). Additionally, we train single-speaker and multi-speaker models using three different text preprocessing pipelines that produce varying degrees of phonetic information and analyze their susceptibility to pronunciation errors. The main contributions of our work can be summarized as follows:

1. We conduct a human evaluation that shows that using externally produced duration information can achieve comparable quality to a student-teacher pipeline.
2. We show that the quality of Estonian grapheme-to-phoneme conversion tools is too low to benefit our use case.
3. Our error analysis shows significant improvements in robustness from using a simple multi-speaker approach. Information about word stress, phonetic quantity, and compound word borders can also be helpful to a smaller extent.

2 Background

2.1 Transformers in TTS

The usage of Transformers (Vaswani et al., 2017) for text-to-speech synthesis was introduced by Li et al. (2019), who used an encoder-decoder model for predicting the frequency distribution of mel-spectrogram frames from English phonemes. Their evaluation of Transformer TTS demonstrated state-of-the-art results. However, the autoregressive nature of predicting each frame sequentially during inference proved too time-consuming to be used in a production setting.

To parallelize inference-time generation, Ren et al. (2019) proposed FastSpeech – a non-autoregressive version of Transformer TTS. The authors used the original autoregressive model as a teacher and extracted character durations from its attention. They trained a student model with a convolutional duration predictor, scaled the encoder outputs to the predicted durations, and generated all output frames in parallel. The mean opinion score (MOS) evaluation on a phonemized English dataset showed no significant decrease in quality when using the non-autoregressive model compared to the much slower Transformer TTS architecture.

FastPitch (Łańcucki, 2021) complemented the duration predictor with a pitch prediction module. The authors used grapheme inputs in their experiments and extracted the duration information from a Tacotron 2 model (Shen et al., 2018). However, they

claimed similar quality when using phonemes and durations from a Montreal Forced Aligner (MFA) model (McAuliffe et al., 2017) – a Kaldi-based (Povey et al., 2011) tool for aligning audio to text. The concurrently developed FastSpeech 2 (Ren et al., 2021) replaced the teacher model with MFA and showed that speech quality improves when conditioning the outputs on both pitch and energy information. Using externally generated durations simplified the training procedure and reduced model development costs. However, this approach assumes that a high-quality alignment model for the language already exists.

2.2 Estonian TTS

While there is existing research on Estonian text-to-speech synthesis, there are limited contributions that use neural methods. There are, however, numerous mentions of using the Vabamorf morphological analyzer (Kaalep and Vaino, 2001) to produce additional phonetic features with other TTS methods (Mihkla et al., 2001; Mihkla, 2007).

A few recent works have included neural text-to-speech to some extent. For example, an analysis by Mihkla (2020) shows that an RNN-based TTS model can generate pronunciations with the correct quantity degree with only 77.8% accuracy. Similarly, convolutional text-to-speech models using grapheme inputs have been shown to produce pronunciation mistakes in up to 17% of all sentences (Rätsep et al., 2020).

At the time of writing, no published works have evaluated the quality of Estonian Transformer-based models or considered using phonemes or other phonetic features with any neural text-to-speech methods. Furthermore, the usability of existing Estonian alignment models, such as the Kaldi-based force alignment model by Alumäe et al. (2018), in non-autoregressive synthesis has not been evaluated.

3 Experimental Setup

3.1 Data

Our experiments used Estonian speech data from 6 male and 4 female speakers. We included the Speech Corpus of Estonian News Sentences (Fishel et al., 2020), which consists of recordings of news articles read by four university students. The remaining speakers are professional actors from the audiobook corpora collected by the Estonian Language Institute (Piits, 2022a,b). Although all actors have also recorded a smaller TTS-specific corpus, initial experiments showed that using these in conjunction with the audiobooks resulted in lower synthesis quality due to differences in speaking styles and recording conditions.

We excluded a subset of 100 sentences per speaker from the training data for evaluation purposes. The rest of the dataset was filtered to remove very long samples (longer than 17.5 seconds). As the news dataset transcriptions are not normalized, and we did not want the potential deficiencies of existing Estonian normalization pipelines to affect model quality, we also removed all sentences requiring normalization. We detected such sentences automatically by comparing the original transcriptions to a normalization script output[1]. The dataset sizes before and after filtering can be seen in Table 1.

[1] `https://github.com/TartuNLP/tts_preprocess_et`

Speaker	Domain	Before filtering		After filtering	
		Samples	Duration (h)	Samples	Duration (h)
Albert (m)	news	12003	22.3	7455	11.6
Indrek (m)	audiobooks	6425	8.6	6108	7.5
Kalev (m)	news	8750	17.6	5273	8.7
Külli (f)	audiobooks	6182	7.5	6017	7.1
Liivika (f)	audiobooks	6564	8.5	6166	7.5
Mari (f)	news	12630	24.1	7641	11.7
Meelis (m)	audiobooks	10140	15.1	9830	14.3
Peeter (m)	audiobooks	5315	7.5	5026	6.4
Tambet (m)	audiobooks	12865	16.9	12410	15.5
Vesta (f)	news	2990	5.7	1910	3.2
Total		83864	133.8	67836	93.6

Table 1: Training dataset sizes for each speaker before and after filtering.

We used three text preprocessing pipelines to create dataset versions with various degrees of phonetic information. As a baseline, we use grapheme inputs. The second pipeline follows the example of existing Estonian TTS research and is based on Vabamorf (Kaalep and Vaino, 2001) to detect word stress, palatalization, phonetic quantity, and compound word borders. The third option uses the Phonemizer library for Python (Bernard and Titeux, 2021) for grapheme-to-phoneme conversion. All dataset versions use lowercased inputs and minimal punctuation normalization rules to reduce the number of symbols in the model vocabulary.

We used the alignment tool by Alumäe et al. (2018) on each training sample to generate alignments between the text and the waveform. This alignment information was used to trim the pauses in the audio files and extract the duration of each grapheme. For Phonemizer and Vabamorf outputs, we created an additional post-processing pipeline to align the input characters with graphemes.

All audio files were resampled at 22050 Hz and converted into mel-scale spectrograms using a raised cosine window (Hann window) with a frame size of 1024 and a hop size of 256 samples.

3.2 Model Configuration

We trained all models using an open-source text-to-speech implementation[2] of the autoregressive Transformer TTS model (Li et al., 2019) and a non-autoregressive model similar to FastPitch (Łańcucki, 2021) that includes explicit character duration and pitch prediction components. For our baseline, we trained grapheme-based Transformer TTS models for each speaker and extracted the duration information from its attention to train the non-autoregressive model. For comparison, we created single-speaker and

[2] `https://github.com/TartuNLP/TransformerTTS`

multi-speaker models with all three text variants using the duration information from the external aligner.

All models were trained for at least 500k steps using a batch size of 12800 frames and identical hyperparameters. The single-speaker and multi-speaker models contained the same number of model parameters, and we did not use additional sampling techniques to mitigate the data imbalance between speakers. The speaker identity in multi-speaker models was controlled by a 2-digit speaker ID prepended to the input text (Wang et al., 2018).

We paired the TTS models with existing HiFiGAN vocoder models[3] (Kong et al., 2020). For speakers Mari, Külli, and Liivika, we used a model trained on the LJ Speech dataset (Ito and Johnson, 2017) and finetuned on ground truth aligned spectrograms produced by Tacotron 2 (Shen et al., 2018). We used a model trained on the VCTK dataset (Yamagishi et al., 2019) for all other speakers.

The appropriate vocoder for each speaker was selected by evaluating TTS samples with both vocoder models. While the LJ Speech vocoder produced excellent speech, it was only suitable for speakers with similar characteristics to the LJ Speech dataset speaker (higher-pitched female speakers). VCTK, however, is a multi-speaker dataset, and the model can be used for a wider variety of speakers. Although a model fine-tuned on our datasets may be optimal in the future, we believe that the use of different vocoders does not have a significant impact on our results as the vocoder selection is consistent between methods, and our objective is not to compare speakers directly to each other.

4 Results

4.1 Speech Quality Evaluation

To evaluate the quality of our models, we conducted a two-part mean opinion score (MOS) evaluation[4] (Chu and Peng, 2001). The first part of our evaluation measured the overall synthesis quality and the effect of using externally produced duration information compared to the student-teacher baseline.

We used a subset of 200 random sentences (20 per speaker) from the held-out dataset to generate evaluation samples for different grapheme-based models. Additionally, the evaluation included original ground truth samples (GT) and versions reconstructed with the HiFiGAN vocoder (GT mel + voc). We used the same subset of sentences to evaluate each approach to ensure comparability between scores, and each sample was evaluated by at least four native Estonian speakers. The evaluation results can be seen in Table 2.

The evaluation results show marginally lower scores for the models using externally produced alignments. Considering the time and infrastructure costs of training teacher models, we argue that the difference in model quality is insignificant and does not justify the student-teacher approach. These results also confirm that the quality of the existing Estonian alignment model is suitable for our use case.

[3] Model files: `https://github.com/jik876/hifi-gan`

[4] Evaluation samples: `https://tartunlp.github.io/TransformerTTS/bhlt2022`

Method	MOS
GT	4.41 ± 0.07
GT mel + voc	4.17 ± 0.07
Baseline	3.47 ± 0.08
Ext. aligner	3.42 ± 0.08
Ext. alginer, multi-speaker	3.44 ± 0.07

Table 2: Mean opinion scores with 95% confidence intervals on the held-out dataset.

We can also see comparable results with the multi-speaker model. While multi-speaker models perform similarly to single-speaker models, they have a significant advantage in production environments where they can synthesize the voice of several different speakers while using the same amount of computational resources.

4.2 Robustness Evaluation

The second part of the evaluation measured the level of robustness when using different degrees of phonetic information. Additionally, we wanted to test whether using multi-speaker models with more training data would decrease the usefulness of this information.

The evaluation included three speakers who represent different data scenarios. Mari and Vesta are examples of high- and low-resource news domain datasets. The third speaker Meelis has plenty of training data, but it contains only one audiobook describing life in 19th century rural Estonia. Therefore, the dataset contains very few examples of words with foreign pronunciation patterns, such as not having the word stress on the first syllable. As a result, we have observed that models trained on this dataset alone are prone to more robustness issues when synthesizing modern texts. Although the evaluation focused on three speakers, it should be noted that datasets from all speakers were used to train the multi-speaker models.

We sampled the evaluation sentences from news and web texts in the Estonian National Corpus (Koppel and Kallas, 2019). We used the same set of 50 evaluation sentences across all speakers and methods and collected at least two evaluations per sample. The results of this evaluation can be seen in Table 3.

The MOS evaluation results show that the evaluators preferred the models that used Vabamorf's features for all three speakers. Additionally, we achieve the best results for high-resource datasets with multi-speaker models. The most significant improvement was for Meelis (+0.68 points), for whom the evaluation sentences are arguably furthest from its training distribution. This supports our hypothesis that the additional training data from other speakers can help reduce robustness issues when synthesizing out-of-domain texts.

The low-resource Vesta dataset is the only speaker that achieves higher results with a single-speaker model. Its MOS is the highest of all three single-speaker models, which suggests that 3.2 hours of training data may be sufficient to achieve comparable quality

Method	Mari	Vesta	Meelis	Average
Grapheme	3.68 ± 0.2	3.83 ± 0.17	3.51 ± 0.24	3.67 ± 0.12
Vabamorf	3.99 ± 0.2	$\mathbf{4.11 \pm 0.17}$	3.52 ± 0.23	3.87 ± 0.12
Phoneme	2.66 ± 0.19	2.92 ± 0.22	2.54 ± 0.23	2.71 ± 0.12
Grapheme, multi-speaker	3.96 ± 0.18	3.84 ± 0.16	3.98 ± 0.18	3.93 ± 0.1
Vabamorf, multi-speaker	$\mathbf{4.04 \pm 0.2}$	3.98 ± 0.16	$\mathbf{4.2 \pm 0.18}$	$\mathbf{4.07 \pm 0.11}$
Phoneme, multi-speaker	2.9 ± 0.21	2.93 ± 0.22	2.63 ± 0.23	2.82 ± 0.13

Table 3: Mean opinion scores with 95% confidence intervals on out-of-domain data.

to models with over 10 hours of data. However, it should be noted that the scores are not directly comparable between different speakers as they are also affected by the vocoder and evaluators' personal preferences. Although Vesta's MOS in multi-speaker models is not significantly lower, the results indicate that the multi-speaker setup should be revised to ensure that the models do not underfit to speakers with less data.

Compared to other models, phoneme-based synthesis achieves significantly lower scores for all speakers. These results suggest that the Estonian grapheme-to-phoneme conversion pipeline quality is insufficient for our use case, and Transformers can learn Estonian pronunciation patterns from graphemes with better results.

In addition to numeric ratings, the evaluators specified the types of pronunciation mistakes that they noticed. These types included word stress errors, incorrect phoneme length, and the use of an incorrect phone. The latter also covered cases of skipping or adding sounds. Additionally, we asked the evaluators to specify whether this error included palatalization mistakes. The error analysis results are provided in Table 4.

Method	Stress	Length	Phone	Palatalization
Grapheme	37.33%	28.67%	9%	1%
Vabamorf	30%	14.33%	13%	2%
Phoneme	58.33%	83.67%	23%	2.67%
Grapheme, multi-speaker	22.67%	14%	**4.33%**	**0.67%**
Vabamorf, multi-speaker	**13%**	**7.67%**	5.33%	1.67%
Phoneme, multi-speaker	59.67%	81.33%	17%	1%

Table 4: Sentence-level pronunciation error rates.

The error analysis demonstrates that the most significant pronunciation improvements always come from using a multi-speaker model. As our multi-speaker setup was quite simple, we believe there is potential for even further improvements if it were revised.

The results also confirm the findings of the MOS evaluation and support the use of Vabamorf's phonetic features. The word stress and phoneme length error rates are lowest with the multi-speaker Vabamorf model. Although improvements over the single-speaker grapheme models are significant, we consider the 13% error rate in word stress placement to be high, and we believe it could have a considerable negative impact on the general perception of Estonian TTS quality.

In terms of palatalization, grapheme inputs achieved marginally better results over Vabamorf, but both approaches yield relatively similar low error rates. Based on this, we believe that the models do not need palatalization information. However, additional analysis with more samples is needed to confirm this. Errors in phone usage follow a similar pattern. This was expected as palatalization is Vabamorf's only annotation type that could have contributed to this category.

Models with phonemized inputs and potentially more information about phone selection perform similarly in the palatalization category. However, the error rates are considerably higher for phone selection, confirming our suspicions about low grapheme-to-phoneme conversion quality.

5 Conclusion

In this work, we evaluated the quality of Estonian text-to-speech synthesis with non-autoregressive Transformer-based models and analyzed whether it can be improved by integrating different language-specific tools. Our experiments showed that Estonian has an existing high-quality alignment model that can be reused for character duration extraction reducing the costs of Estonian TTS development in the future. Additionally, our analysis suggests that while Estonian phoneme-to-grapheme conversion quality is low, synthesis still benefits from using annotations that mark word stress, quantity degrees, and compound word borders.

Our main takeaway, however, is the importance of using more data by training multi-speaker models. Multi-speaker models are not only more versatile in production environments but we also found them to be more robust and less susceptible to pronunciation errors. Therefore, we believe that future Estonian TTS research should focus on multi-speaker models and look into more options for leveraging additional speech data that would also be applicable to other languages with no existing phonetic annotation tools.

References

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

Bernard, M., Titeux, H. (2021). Phonemizer: Text to phones transcription for multiple languages in Python, *Journal of Open Source Software* **6**(68), 3958.

Chu, M., Peng, H. (2001). An objective measure for estimating MOS of synthesized speech, *EUROSPEECH 2001, 7th European Conference on Speech Communication*, ISCA, pp. 2087–2090.

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

Ito, K., Johnson, L. (2017). The LJ Speech dataset, https://keithito.com/LJ-Speech-Dataset/.

Kaalep, H.-J., Vaino, T. (2001). Complete morphological analysis in the linguist's toolbox, *Congressus Nonus Internationalis Fenno-Ugristarum Pars V* pp. 9–16.

Kong, J., Kim, J., Bae, J. (2020). HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis, *Advances in Neural Information Processing Systems*, Curran Associates, Inc., pp. 17022–17033.

Koppel, K., Kallas, J. (2019). Estonian national corpus, https://doi.org/10.15155/3-00-0000-0000-0000-08565L.

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

Li, N., Liu, S., Liu, Y., Zhao, S., Liu, M. (2019). Neural speech synthesis with Transformer network, *Proceedings of the AAAI Conference on Artificial Intelligence*, Association for the Advancement of Artificial Intelligence, pp. 6706–6713.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi, *Proc. Interspeech 2017*, pp. 498–502.

Mihkla, M. (2007). Modelling speech temporal structure for Estonian text-to-speech synthesis: Feature selection, *Trames* **11**(3), 284–298.

Mihkla, M. (2020). Väldete analüüs sünteesi teel, *Keel ja Kirjandus* **63**(11), 935–950.

Mihkla, M., Meister, E., Kiissel, I., Lasn, J. (2001). Evaluation the quality of Estonian text-to-speech synthesis and diphone corrector for the TTS system, *Dialogue '2001: Computational Linguistics and its Applications: International Workshop, Proceedings. Vol. 2. Applications*, pp. 385–390.

Mihkla, M., Piits, L., Nurk, T., Kiissel, I. (2008). Development of a unit selection TTS system for Estonian, *Proceedings of the Third Baltic Conference on Human Language Technologies*, IOS Press, pp. 181–187.

Nurk, T. o. (2012). Creation of HMM-based speech model for Estonian text-to-speech synthesis, *Human Language Technologies - the Baltic Perspective: Proceedings of the Fifth International Conference*, IOS Press, pp. 162–168.

Piits, L. (2022a). Estonian female voice audiobook corpus for speech synthesis, https://doi.org/10.15155/3-00-0000-0000-0000-090D4L.

Piits, L. (2022b). Estonian male voice audiobook corpus for speech synthesis, https://doi.org/10.15155/3-00-0000-0000-0000-08BF4L.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K. (2011). The Kaldi speech recognition toolkit, *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.

Rätsep, L., Piits, L., Pajupuu, H., Hein, I., Fishel, M. (2020). Neural speech synthesis for Estonian, *arXiv preprint arXiv:2010.02636*.

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

Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., Liu, T.-Y. (2019). FastSpeech: Fast, robust and controllable text to speech, *Advances in Neural Information Processing Systems*, Curran Associates, Inc.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R. J., Saurous, R. A., Agiomyrgiannakis, Y., Wu, Y. (2018). Natural TTS

synthesis by conditioning WaveNet on mel spectrogram predictions, *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4779–4783.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., Polosukhin, I. (2017). Attention is all you need, *Advances in Neural Information Processing Systems*, Curran Associates, Inc.

Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R. J., Battenberg, E., Shor, J., Xiao, Y., Ren, F., Jia, Y., Saurous, R. A. (2018). Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis, *arXiv preprint arXiv:1803.09017*.

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

Received August 19, 2022 , accepted August 27, 2022