BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2023
http://excel.fit.vutbr.cz

Text-to-speech Personalization

Michal Luner

Abstract

This work aims to develop a model that can convert Czech input text into speech that closely resembles a target speaker. The chosen approach involves training a base text-to-speech model using a large sample of data and fine-tuning it on a smaller personalized dataset for a specific speaker, with improvements monitored using objective and subjective metrics. Several fine-tuned models were developed, with the male model achieving a mean opinion score of 4.12/5 (4.59/5 for the ground truth samples). The work contributes to the development of high-quality Czech TTS models and datasets by describing the necessary steps and techniques.

*xluner01@vut.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

Creating personalized text-to-speech (TTS) models with limited training data is a challenging task, especially with the lack of publicly available high-quality Czech TTS datasets. This work aims to contribute to the development of Czech TTS models and datasets by outlining the necessary steps and techniques to train a high-quality TTS model.

Training a TTS model requires higher-quality audio samples than an automatic speech recognition (ASR) model. However, due to the limited availability of TTS training data, this work utilizes one of the best available Czech ASR datasets to train a base TTS model. The goal is to develop a personalized model that generates audio samples, which closely resemble the target speaker, with low error rates (such as EER, WER, and CER). The final model should also perform well in subjective listening tests.

The existing TTS solutions for the Czech language are limited, with most of the Czech models and datasets being proprietary. In this project, the VITS [1] model, which achieved a mean opinion score (MOS) of 4.43/5[1] in English on the LJ Speech dataset, is utilized.

The chosen approach consists of training a base TTS model using a large dataset and fine-tuning it on a smaller, personalized dataset of a specific speaker. To track the progress between the models, an objective evaluation pipeline is deployed on each one.

The contributions of this work include developing a TTS model that generates high-quality speech for a specific target speaker in Czech using limited training data resources. The project also provides a pipeline for evaluating generated samples based on objective metrics and another pipeline for dataset creation that can be used for any language.

2. Poster Description

2.1 TTS model

Figure 1 describes the idea of a general TTS model. An input sequence is passed to the model, which outputs a waveform.

The VITS [1] architecture, depicted inside of the blue arrow, outlines the steps required for audio generation. The text input is first converted into phonemes, which are fed into the text encoder. Using a stochastic duration predictor, the duration of each phoneme in the input sequence is computed. This duration alignment is further passed to the normalizing flow, which transforms it into a latent representation that the decoder can understand. The decoder is trained to output the final waveform. In addition, the audio generation can be conditioned on a speaker embedding, which uniquely identifies the speaker's vocal characteristics.

The training of the VITS model consists of passing the corresponding text/audio pairs. An additional encoder processes the input spectrogram[2] and converts it into

[1] With MOS of 4.46/5 for the ground truth samples

[2] Spectrogram represents the ground truth audio sample.

the latent space. The text encoder does the same thing with the input text. By deploying several loss functions, the ultimate goal is to train the model to generate latent representation of the resulting audio sample just on the input text.

2.2 Datasets

To train a base model, subset of data from the Par-Czech dataset [2] was used. This dataset comprises parliamentary hearings, which is probably the best publicly available Czech source of a large amount of data. Once the model performed well based on objective metrics, personalized datasets[3] were developed for male and female speakers.

In light of recent political events, Petr Pavel and Danuše Nerudová, who have given numerous studio interviews, were chosen as the speakers.

As shown in Figure 2, the first step was to separate the interviews into chunks of speech, where the targeted speaker (orange) speaks.

However, these chunks could be several minutes long, therefore, the audio files were further segmented.

Finally, the dataset was manually filtered. This consisted of reviewing each audio file with its transcription and correcting possible typos that might have occurred during the transcription process.

2.3 Evaluation

Figure 3 shows the improvement of the base and fine-tuned model. The plot represents a distribution of cosine distance results between speaker embeddings, where a lower value indicates better reproduction of speech characteristics.

The final fine-tuned models were evaluated using listening tests mainly focusing on speech naturalness and intelligibility. The male model achieved MOS of 4.12/5[4] and the female model scored MOS of 3.05/5[5]. Although the male model performed well, the female model did not improve its performance, even though it was fine-tuned on female speakers before personalization.

Moreover, the models were fine-tuned using a limited amount of data to evaluate the minimum amount required for training a well-performing personalized model. Surprisingly, the models trained on as little as 10 minutes of data performed almost as well as the original fine-tuned models that used over an hour of training data.

[3] The entire pipeline for dataset generation reflects the desired properties of a TTS dataset as described in LibriTTS paper [3].

[4] With MOS of 4.59 for the ground truth samples.

[5] With MOS of 4.55 for the ground truth samples.

3. Conclusions

The findings of this work demonstrate that it is indeed possible to train a personalized Czech TTS model using limited training data. Furthermore, it proposes a semi-automated solution for creating new datasets, which could contribute to the development of new models.

For future works, exploring options for different female speakers and fine-tuning the personalized model on even cleaner data could be potential areas of research.

Acknowledgements

I would like to thank my supervisor Ing. Jan Brukner for guiding me and helping me with all the problems I have encountered.

References

[1] Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pages 5530–5540. PMLR, 2021.

[2] Matyáš Kopp, Vladislav Stankov, Jan Oldřich Kruza, Pavel Straňák, and Ondřej Bojar. Par-czech 3.0: A large czech speech corpus with rich metadata. In *International Conference on Text, Speech, and Dialogue*, pages 293–304. Springer, 2021.

[3] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*, 2019.