BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2026
http://excel.fit.vutbr.cz

Automation of Data Collection for Speech Synthesis Training

Michal Luner

Abstract

Developing high-quality, expressive Text-to-Speech (TTS) systems requires vast amounts of annotated audio data, which remains scarce for the middle-resourced languages, such as Czech. To bridge this gap, this work introduces a scalable, automated pipeline that utilizes "in-the-wild" audio from YouTube and processes it using source separation, automatic speech recognition (ASR), and semantic segmentation. Out of 25.4k hours of raw audio, the pipeline successfully distilled 9.7k hours of clean, diverse speech data, yielding a retention rate of 38.4%. This extensive dataset, validated by training models such as VITS and F5-TTS, lowers the barrier to future research and development of expressive Czech speech synthesis.

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

Introduction

The quality, naturalness, and expressivity of modern Text-to-Speech (TTS) models are directly bottlenecked by their training data. While English benefits from massive, open-source datasets, the Czech language suffers from a lack of publicly available, large-scale TTS data.

Collecting and annotating thousands of hours of studio-quality speech manually is expensive and time-consuming. The core challenge is to design an automated system capable of scraping "in-the-wild" internet audio and autonomously filtering out noise, overlapping speech, and bad transcriptions to output a dataset clean enough for strict TTS training constraints.

Recently, automated pipelines for annotated speech data collection, such as the Emilia-Pipe pipeline [1], have been introduced. While these provide a strong baseline, they are primarily optimized for high-resource languages. When applied to Czech, they suffer from language-specific ASR hallucinations, suboptimal chunking mechanisms, and performance bottlenecks, resulting in mediocre data quality.

An optimized and parallelized data collection pipeline tailored for the Czech language is proposed. Semantic-level splitting, rather than naive duration-based cuts, is introduced, and a compression-ratio filtering mechanism is employed to eliminate ASR hallucinations.

A massive, automatically curated Czech speech dataset totaling 9.7k hours of clean audio is produced. In addition, an optimized, open-source data processing pipeline is provided, and the quality of the curated data is validated through the successful training and evaluation of two speech synthesis models.

1. Developed Pipeline

As illustrated in Figure 1, the core of this work is the developed automated pipeline. The process begins with raw "in-the-wild" audio, primarily sourced from YouTube videos.

Because YouTube audio might contain background music, sound effects, and multiple speakers, the **Preprocessing** module first standardizes the audio and applies robust Source Separation to isolate human vocals. Next, using Voice Activity Detection and Speaker Diarization in the **Segmentation** and **Merge**, silences and overlapping speakers are identified for removal, and new audio segments are created. The architecture of the modules ensures that sentences are not broken mid-word or mid-thought, preserving the natural prosody required for expressive TTS.

Following the **LID** module for language identification of each segment, the vocals are transcribed using NVIDIA Canary inside the **ASR** module.

Once transcribed, the **Filtering** block evaluates the audio quality using **DNSMOS** (a neural network-based

mean opinion score estimator), screens for ASR hallucinations, and performs other filtering.

Finally, a **Clean Dataset** is obtained, consisting of audio segments, transcriptions, and optionally speaker embeddings.

2. Processed Data & Challenges

Table 1 summarizes the scale of the data processed during this project. The initial raw data consisted of 25.4k hours sourced from 100 diverse Czech YouTube channels, intentionally chosen to capture a wide range of speakers, dialects, and domains.

After passing through the pipeline, 9.7k hours of clean speech were retained. This 38.4% retention rate reflects the strictness of the filtering modules. It ensures that only intelligible speech is included in the final dataset.

During development, several key challenges had to be overcome:

- Text hallucinations produced by ASR models were successfully mitigated through the implementation of a filter based on the compression ratio.
- The processing of 25k hours of audio is computationally demanding. Extensive pipeline parallelization was therefore required to enable processing within a feasible timeframe on the university cluster.

3. Text-to-Speech Models & Results

To validate the pipeline's output, two distinct Text-to-Speech architectures were trained:

- **VITS** [2]: A representative of a classic end-to-end TTS model.
- **F5-TTS** [3]: A modern, efficient Flow-Matching architecture built on a Diffusion Transformer.

As shown in Table 2, the models are evaluated across multiple objective metrics, including Word Error Rate (WER), Speaker Similarity (SIM), and overall audio quality (DNSMOS). A comparison between models trained on the developed ML-Pipe dataset and those trained on the existing ParCzech4 dataset shows that the proposed data provides strong performance and, in several cases, surpasses the existing baseline. Furthermore, a demo webpage[1] has been established to allow listeners to subjectively evaluate the generated audio samples.

[1] A QR code will be attached next to the poster.

Conclusions

The lack of Czech TTS data is overcome by the modular ML-Pipe pipeline, which is designed for easy adaptation to other low-resource languages. The main outcomes of this work are the processing pipeline itself, a 9.7k-hour dataset, and models capable of generating high-quality speech.

While ASR error correction was not included in this study, it is identified as a valuable field for future exploration.

Acknowledgements

I would like to sincerely thank my supervisor, Ing. Jan Brukner, for his continuous guidance, technical insights, and support throughout the development of this pipeline and thesis.

Special thanks go to Ing. Igor Szőke, Ph.D., for providing the source data used in this work, and to Ing. Karel Beneš, Ph.D., for his guidance and support in ASR.

References

[1] Haorui He, Ziyue Shang, Chaohong Wang, Xingchen Li, et al. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. arXiv preprint, 2024. https://arxiv.org/abs/2407.05361.

[2] 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 (ICML)*, 2021.

[3] Yushen Chen, Zhikang Nin, Ziyang Ma, Keqi Deng, et al. F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching. arXiv preprint, 2024. https://arxiv.org/abs/2410.06885.