Automation of Data Collection for Speech Synthesis Training

Author: **Bc. Michal Luner**
xluner01@vutbr.cz

Supervisor: **Ing. Jan Brukner**
ibrukner@fit.vutbr.cz

Motivation & Main Ideas

- Lack of publicly available Czech TTS datasets.
- Analysis of existing automated pipelines for annotated speech data collection [He et al., 2024].
- Evaluation & optimization for the Czech language.
- Validation of the curated data by evaluating TTS.

Processed Data

Table 1

Status	Total
Raw	25.4k h
Processed	**9.7k h**
Retention	**38.4 %**

- Sourced from 100 Czech YouTube channels.
- High diversity in speakers, dialects, and domains.

Challenges

- ASR hallucinations → filter by compression ratio.
- Semantic splitting rather than duration-based.
- Performance → pipeline parallelization.

Text-to-Speech Models

- **VITS** [Kim et al., 2021]: A representative of classic end-to-end TTS.
- **F5-TTS** [Chen et al., 2024]: Modern, efficient Flow-Matching architecture, built on a Diffusion Transformer.

Project Outcomes & Results

- A source-language-independent data processing pipeline for NLP tasks.
- Several TTS models trained on both developed and public datasets.
- A demo webpage featuring generated audio samples.

Table 2: Evaluation of TTS trained on different Czech datasets. WER [%], SIM ∈ [−1, 1], and DNSMOS ∈ [1, 5].

Model	Dataset	WER ↓	DNSMOS ↑	SIM ↑
F5-TTS	ML-Pipe	**7.26**	**3.16**	**0.85**
	ParCzech4	13.70	3.12	0.75

Developed Pipeline

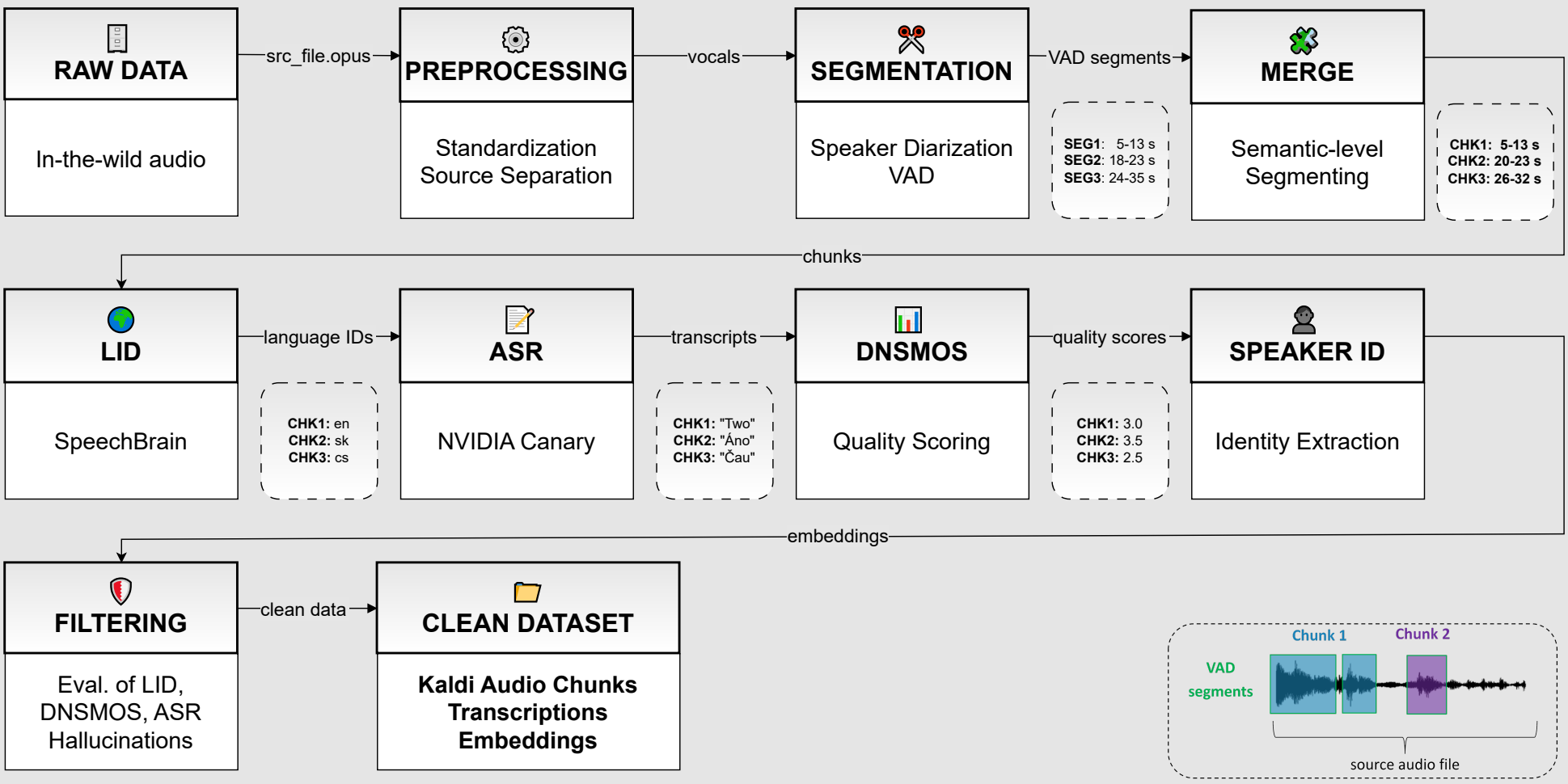

Figure 1: Proposed pipeline architecture tailored for Czech speech data processing.

References

Chen, Y., Niu, Z., Ma, Z., Deng, K., Wang, C., Zhao, J., Yu, K., and Chen, X. (2024). F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*.

He, H., Shang, Z., Wang, C., Li, X., Gu, Y., Hua, H., Liu, L., Yang, C., Li, J., Shi, P., Wang, Y., Chen, K., Zhang, P., and Wu, Z. (2024). Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *Proc. of SLT*.

Kim, J., Kong, J., and Son, J. (2021). Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pages 5530–5540. PMLR.