Text-to-speech Personalization

Author: Michal Luner
Supervisor: Ing. Jan Brukner

Motivation and goals

- Lack of available high-quality text-to-speech models for the Czech language.
- Exploring automatic speech recognition datasets and using them for text-to-speech system training.
- Developing a personalized model producing high-quality audio samples closely resembling the target speaker.

Input sentence

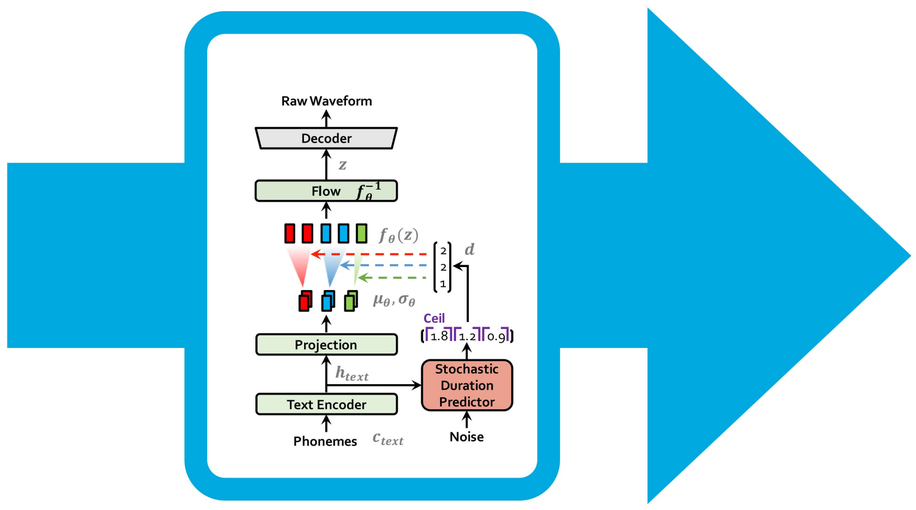
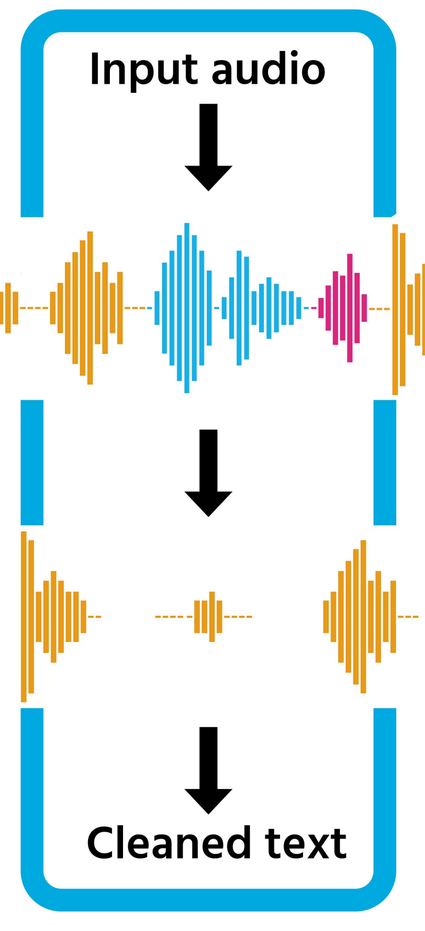


Figure 1

Solution

- Parlamentary hearings were used to train a base text-to-speech model.
- The base model was fine-tuned using a created dataset.
- The generated audio samples were evaluated using objective and subjective metrics.

Figure 2



Results

- Two personalized models achieving naturalness of 4.12/5 (original data 4.59) were trained with as little as 10 minutes of target speaker data.
- Evaluation pipeline for the generated audio samples was developed.
- Semi-automatic dataset development pipeline was created.
- Two datasets, one male and one female, were produced.

Figure 3