Data Selection Strategies for Multi-Speaker TTS in Lithuanian

Progress Report 2

Aleksandr Jan Smoliakov[1]

Supervisor: Gerda Ana Melnik-Leroy
Co-Advisor: Gražina Korvel

[1]Vilnius University, Faculty of Mathematics and Informatics

2025–11–19

Shift in research focus

The research focus has changed.

- **Previous:** Focus on text normalization strategies for TTS.
- **Current:** Training robust multi-speaker TTS models.

Motivation for change:

- Discovery of the *Liepa 2* dataset.
- *Liepa 2* provides significantly larger and higher-quality annotated data than *Common Voice*.
- The transcriptions are pre-normalized, thus normalization is no longer a concern.

Liepa 2 Dataset

Dataset overview:

- **1000 hours** of Lithuanian speech.
- **2621** distinct speakers.
- Varied recording conditions (Studio, Radio, TV, etc.).

The challenge:

- The “top” speaker has only **2.5 hours** of speech.
- High-quality single-speaker TTS usually requires at least **4 to 20 hours** of speech.
- Implication: We cannot train a high-quality model on a single speaker alone.

Solution: Train multi-speaker models to leverage shared linguistic features.

Feasibility

Computational constraints:

- Training on the full 1000h corpus is computationally expensive and time-consuming.
- Limited GPU resources necessitate smaller training sets.
- We need strategies to select effective subsets of data.

Research Questions:

1. How should we select subsets of data for training multi-speaker TTS models?
2. How does the number of speakers in the training data affect the quality of synthesized speech?

Methodology

Models:

- **Tacotron 2 Variants:**
    - Using DCA (Dynamic Convolution Attention) and DDC (Double Decoder Consistency).
    - Adapted for pre-trained multi-speaker (VCTK Corpus) HiFi-GAN v2 vocoder.
- **VITS:** End-to-end pipeline (Glow-TTS + HiFi-GAN).

Subset selection strategies:

- **Top-N speakers:** Selecting only speakers with the most data.
- **Random sampling:** Selecting utterances uniformly at random.
- **Balanced sampling:** Optimizing for speaker diversity (preferring more speakers with fewer utterances each).

Progress summary

Completed:

- ✓ Data preprocessing pipeline
- ✓ Speaker embeddings computation
- ✓ Text accentuation
- ✓ Solved Tacotron 2 technical issues
- ✓ Tacotron 2 adaptation for HiFi-GAN
- ✓ Initial Tacotron 2 experiments

Ongoing:

- VITS configuration
- Implementing subset selection strategies
- Training final multi-speaker TTS models
- Quality evaluation

Next Steps:

1. Run experiments with defined subsets.
2. Analyze effect of speaker count on quality.
3. Write Thesis.

Example of generated output

socialinės savidestrukcijos spiralės multiplikavimas technologinėmis priemonėmis yra dvidešimt pirmo amžiaus fenomenas

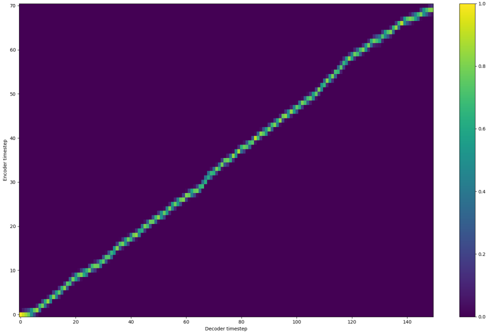

Figure: Test sample's input-output alignment from Tacotron 2-DCA model

Speaker IS031: Click to play

Speaker MS003: Click to play

Thank You!

Thank you for your attention!