Training Data Selection Strategies for Multi-Speaker TTS in Lithuanian

Pre-defense Presentation

Aleksandr Jan Smoliakov[1]

Supervisor: Dr. Gerda Ana Melnik-Leroy
Scientific Advisor: Dr. Gražina Korvel

[1]Vilnius University, Faculty of Mathematics and Informatics

2025–12–22

Research aim & novelty

Problem statement: Multi-speaker TTS systems have substantial data requirements, which is challenging for low-resource languages.
Data availability: *Liepa 2* corpus offers 939 hours of Lithuanian speech, but individual speaker data is sparse (avg: 21 min/speaker).

Research aim

Measure how varying training dataset **breadth** (number of speakers) and **depth** (duration per speaker) affects the synthesis quality of multi-speaker TTS models under a fixed data budget.

Novelty of the work:

- First systematic study of "breadth vs. depth" trade-offs specifically for the Lithuanian language.
- Analysis of AR (Tacotron 2) vs. NAR (Glow-TTS) models in low-depth settings (7.5 min/speaker).

Methodology: Constant "data budget"

Three subsets of the *Liepa 2* corpus are created. To ensure fair comparison, the total training data budget is fixed at **22.5 hours** for all experiments.

1. Depth	**2. Balance**	**3. Breadth**
30 speakers $\times$ 45 min/speaker	60 speakers $\times$ 22.5 min/speaker	180 speakers $\times$ 7.5 min/speaker

* Speakers are nested ($30 \subset 60 \subset 180$) and gender-balanced.

Methodology: Acoustic models

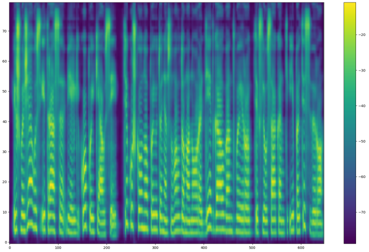
Two distinct architectures were trained from scratch on all three data subsets (6 models total).

- **Tacotron 2:**
  - Autoregressive sequence-to-sequence.
  - High naturalness but slower inference.
- **Glow-TTS:**
  - Non-autoregressive, flow-based.
  - Parallel generation, faster inference.

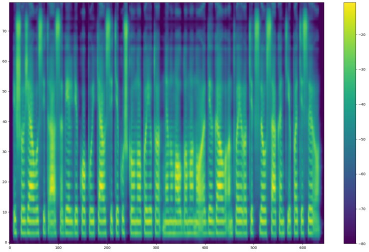
Vocoder: *HiFi-GAN* (pre-trained on VCTK) used for all synthesis to isolate acoustic model performance.

Results: Objective metrics

- All models converged successfully with clear alignment patterns.
- Reducing data per speaker (from 45 min to 7.5 min) had **minimal impact on objective metrics (MCD, F_0 RMSE)**.
- Tacotron 2 produces more dynamic pitch contours.



(a) Tacotron 2 (60 speakers)



(b) Glow-TTS (60 speakers)

Figure: Mel-spectrograms generated by Tacotron 2 (left) and Glow-TTS (right) for the same input speaker and text.

Results: Subjective evaluation (MOS)

21 native Lithuanian listeners evaluated the naturalness of 6 speakers' synthesized speech.

Table: Mean Opinion Score (MOS, 1–5 scale). Higher is better.

Trainset composition	**Tacotron 2**	**Glow-TTS**
30 spk. $\times$ 45 min	3.11 ± 0.16	2.13 ± 0.12
60 spk. $\times$ 22.5 min	$\mathbf{3.12} \pm \mathbf{0.17}$	2.17 ± 0.15
180 spk. $\times$ 7.5 min	3.03 ± 0.18	2.02 ± 0.14

Takeaways:

- In terms of speech naturalness, **Tacotron 2** significantly outperforms **Glow-TTS** across all data selection strategies.
- There are **no significant MOS differences** between the three data selection strategies for either model.

Results: MOS by speaker

Table: Average MOS per speaker across all models.

Speaker ID	Tacotron 2	Glow-TTS
AS009	**4.17 ± 0.20**	**2.61 ± 0.23**
IS031	3.26 ± 0.20	2.13 ± 0.19
IS038	3.48 ± 0.21	2.50 ± 0.19
MS052	2.26 ± 0.17	1.87 ± 0.16
VP131	2.43 ± 0.19	1.93 ± 0.16
VP427	2.92 ± 0.22	1.59 ± 0.14

Takeaways:

- Certain speakers consistently yield higher/lower MOS across models, indicating inherent speaker characteristics may impact synthesis quality.

Remaining work

1. **Writing final thesis:**
   - Corrections based on supervisor feedback.
   - Post-investigation of speaker-wise differences in MOS.
   - Finalizing the "Results", "Discussion" chapters.
   - Plots and tables, formatting, proofreading.
2. **Cleaning up the TTS codebase** for publishing in the GitHub repository
3. Preparing slides, audio samples for the **final defense presentation**.

Thank You!

Thank you for your attention!