Data Selection Strategies for Multi-Speaker TTS in Lithuanian

Progress Report 3

Aleksandr Jan Smoliakov[1]

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

[1] Vilnius University, Faculty of Mathematics and Informatics

2025–12–16

Recap & Experimental design

Liepa 2 challenge:

- 1000 hours of audio, distributed across 2,621 speakers.
- Most speakers have < 30 minutes of data.
- Standard TTS requires 10–20 hours single-speaker data.
- **Core RQ**: Under a fixed data budget, what is the optimal data selection strategy for Lithuanian multi-speaker TTS?

Three datasets, fixed trainset budget (22.5 h audio)

- **Depth**: 30 speakers, 45 min/speaker
- **Balance**: 60 speakers*, 22.5 min/speaker
- **Breadth**: 180 speakers*, 7.5 min/speaker

*Speakers are nested ($30 \subset 60 \subset 180$) and gender-balanced (50/50).

Model architectures

Two distinct acoustic model architectures were trained on all three subsets (6 experiments total).

Tacotron 2

- Autoregressive sequence-to-sequence model.
- Trained for 200 epochs (on each subset) — until convergence.

Glow-TTS

- Flow-based generative model.
- Trained for 400 epochs (on each subset) — until convergence.

Vocoder: Pre-trained **HiFi-GAN v2** (frozen) used for all models.

Objective results

Metrics evaluated on held-out test set (60 sentences).

Model	Speakers (N)	MCD (dB)	F0 RMSE (Hz)
Tacotron 2	30	9.58	31.28
	60	**9.55**	**30.49**
	180	9.63	31.06
Glow-TTS	**30**	**9.90**	37.86
	60	10.00	36.18
	180	9.98	**35.69**

Table: Tacotron 2 moderately, but consistently outperforms Glow-TTS in spectral and pitch accuracy (lower is better).

Observation: Both models were surprisingly insensitive to data composition in terms of objective metrics.

Subjective evaluation (MOS)

Methodology:

- 60 sentences (being) rated by 21 Native Lithuanian speakers.
- Latin Square Design to mitigate bias.
- Scale: 1 (Bad) to 5 (Excellent).
- Tacotron 2 significantly outperforms Glow-TTS in naturalness.

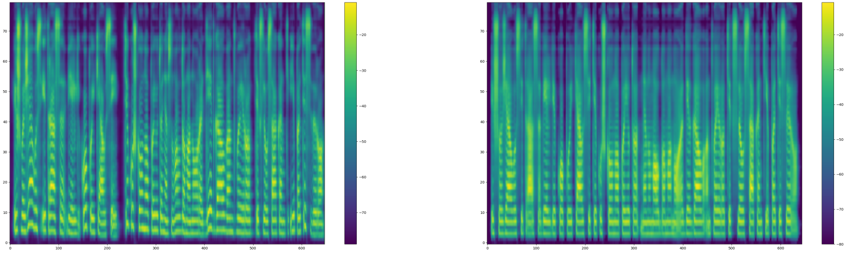
Configuration	Tacotron 2 MOS	Glow-TTS MOS
30 Speakers (Depth)	3.39	2.40
60 Speakers (Balance)	**3.48**	2.39
180 Speakers (Breadth)	3.27	2.16

MOS by speaker

Table: Average MOS per speaker across all models.

Speaker ID	**Tacotron 2 MOS**	**Glow-TTS MOS**
AS009	**4.22**	**2.68**
IS031	3.28	2.28
IS038	3.60	2.58
MS052	2.19	1.94
VP131	2.54	2.00
VP427	3.18	1.64

Visual analysis



(a) Tacotron 2 (60 speakers)

(b) Glow-TTS (60 speakers)

Figure: Mel-spectrograms generated by Tacotron 2 and Glow-TTS for the same input text.

Spectrogram comparison: Tacotron 2 generates finer spectral details and more dynamic pitch contours compared to the “flatter” output of Glow-TTS.

Discussion & Conclusions

- Tacotron 2 consistently produced more natural speech than Glow-TTS across all data strategies.
- Overall, data composition had a limited effect on objective metrics for both models, and only a moderate effect on subjective naturalness.
- All models' synthesis quality strongly depended on individual speaker characteristics.
- Using 7.5 minutes per speaker (180 speakers) is viable for intelligible multi-speaker TTS in low-resource settings.

Progress summary

Completed:

- ✓ Trained 6 multi-speaker TTS models
- ✓ Conducted objective evaluations
- ✓ Preliminary subjective results and drafted findings

Ongoing:

- Subjective evaluation (MOS study)
- Analysis of subjective results
- Writing thesis draft

Thank You!

Thank you for your attention!