Training Data Selection Strategies for Multi-Speaker Text-to-Speech Synthesis in Lithuanian

Master's Thesis Defense

Aleksandr Jan Smoliakov[1]

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

[1]Vilnius University, Faculty of Mathematics and Informatics

2026–01–16

Context & Problem Statement

- **Text-to-Speech (TTS)** systems (also known as speech synthesis) are widely used in virtual assistants, visual aids, and other applications.
- State-of-the-art **Neural TTS** typically requires 10–20 hours of high-quality *single-speaker* data.
- **Low-resource languages** like Lithuanian rarely possess such datasets.
- *Liepa-2* corpus contains 939 hours of annotated Lithuanian speech, but it is **fragmented** across 2,621 speakers.
- **Multi-speaker TTS** models can leverage such data.
- **Question:** What is the optimal strategy for composing multi-speaker datasets to maximize synthesis quality?

Aim & Hypothesis

Research Aim

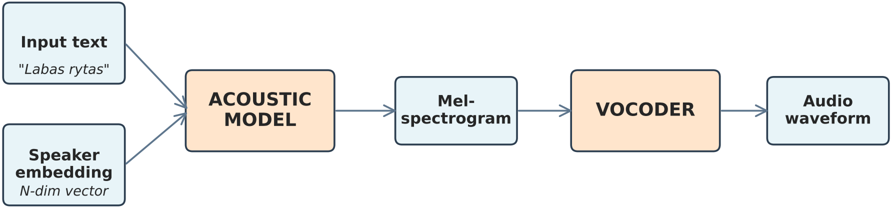
Investigate how varying training dataset **breadth** (number of speakers) and **depth** (duration per speaker) affects the synthesis quality of multi-speaker TTS models under a constant total data budget.

Hypothesis: Data *depth* is the critical factor. Synthesis quality will degrade as the number of speakers increases with the total training duration held constant.

Research Objectives

1. Prepare 3 subsets of *Liepa-2* with varied breadth/depth but constant total duration.
2. Configure and train two distinct acoustic model architectures — one autoregressive (AR), one non-autoregressive (NAR).
3. Evaluate the synthesis quality of the trained models using established objective evaluation metrics.
4. Develop a subjective evaluation application and conduct listening tests to assess naturalness.

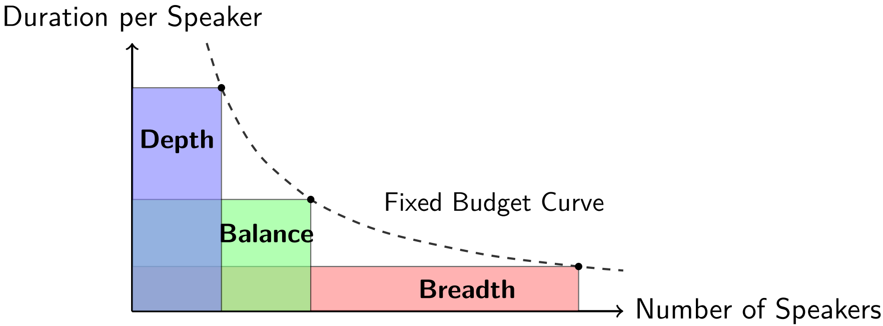
TTS Synthesis Pipeline



TTS Synthesis Pipeline

A typical neural TTS pipeline. An **Acoustic model** generates Mel-spectrograms from input text, which are then converted to raw audio waveforms by a **Vocoder**.

Experimental Design: Constant “Data Budget”



Total “data budget” is fixed. The three strategies represent different points along the breadth-depth trade-off curve.

Experimental Design: Constant "Data Budget"

Total "data budget" fixed at **22.5 hours**.
Speakers are nested ($30 \subset 60 \subset 180$) and gender-balanced to ensure fair comparison.

Strategy	**Speakers**	**Depth/Speaker**	**Total Budget**
Depth	30	45.0 min	22.5 h
Balance	60	22.5 min	22.5 h
Breadth	180	7.5 min	22.5 h

Data Preparation

- **Source:** *Liepa-2* corpus — 939 hours of read Lithuanian speech from 2,621 speakers.
- **Segmentation:** Utterance-level audio segments sliced using provided timestamps.
- **Filtering:** Adult (18+ years) speakers only, read speech (not spontaneous).
- **Text:** Grapheme normalization done using rule-based preprocessing. Kirčiuoklis-based accentuation applied (where accents are unambiguous, 82% of total words).
- **Audio:** Resampled to 22.05 kHz to match pre-trained vocoder.

Model Architectures

Tacotron 2

- Autoregressive.
- Seq2seq (encoder-decoder) architecture.
- Uses Dynamic Convolution Attention.

Glow-TTS

- Non-autoregressive (parallel).
- Flow-based generative model.
- Uses Monotonic Alignment Search.

Speaker Embeddings: 512-dimensional learnable embeddings, jointly trained with the acoustic models.

Vocoder: *HiFi-GAN* (pre-trained on VCTK). Frozen during training to isolate acoustic model performance.

Model Training

- **Hardware:** Personal high-performance workstation with a 48-core CPU, 256 GB of RAM, and an NVIDIA RTX 3090 GPU.
- **Hyperparameters:** Based on default configurations, with adjustments to Mel-spectrogram parameters (to match vocoder), batch size (due to hardware constraints), learning loss schedules (for a balance between convergence speed and optimal loss values), Lithuanian grapheme set, and other minor tweaks empirically found to reduce validation loss.
- **Training:** On each dataset, from scratch, until validation loss convergence.
  - Tacotron 2 for 90k steps ($\approx$ 22 hours) each.
  - Glow-TTS for 180k steps ($\approx$ 25 hours) each.

Evaluation Setup

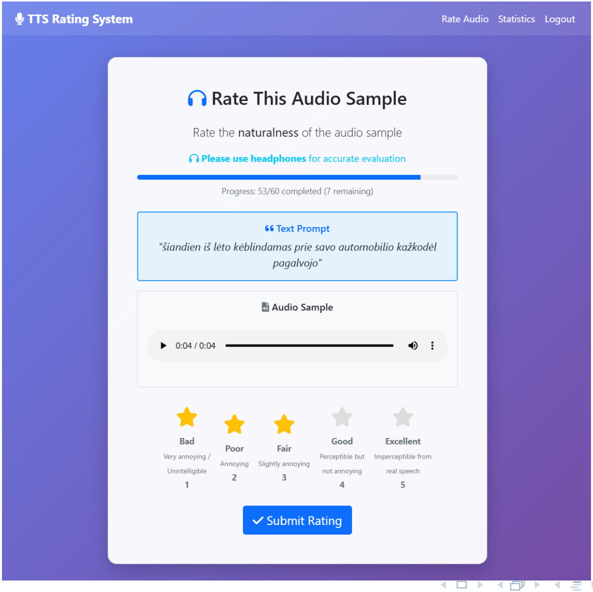
Objective Metrics:

- Mel-Cepstral Distortion (MCD): measures spectral distortion in dB.
- Fundamental Frequency RMSE (F_0 RMSE): measures pitch contour error in Hz.

Subjective Evaluation:

- Mean Opinion Score (MOS) test
- 6 randomly selected speakers
- 60 identical test sentences for each speaker
- Latin Square design

Evaluation Setup: Custom MOS Application



TTS Rating System

Rate Audio | Statistics | Logout

Rate This Audio Sample

Rate the **naturalness** of the audio sample

Please use headphones for accurate evaluation

Progress: 53/60 completed (7 remaining)

Text Prompt

"šiandien iš lėto kėblindamas prie savo automobilio kažkodėl pagalvojo"

Audio Sample

▶ 0:04 / 0:04

★ **Bad** Very annoying / Unintelligible 1

★ **Poor** Annoying 2

★ **Fair** Slightly annoying 3

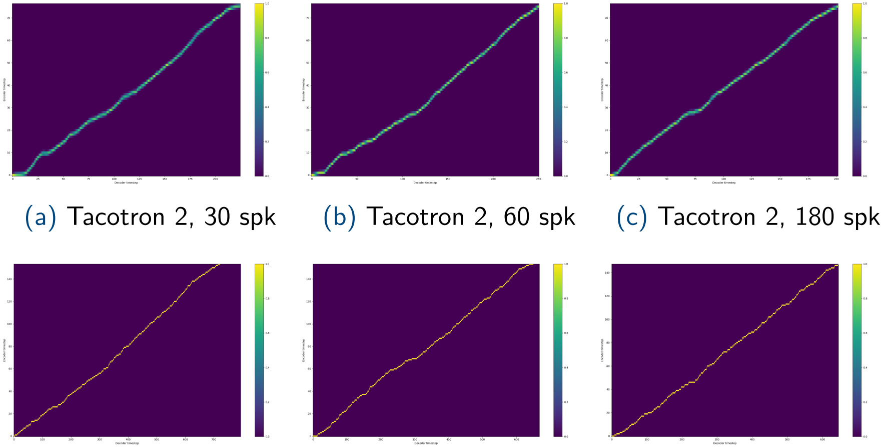
☆ **Good** Perceptible but not annoying 4

☆ **Excellent** Imperceptible from real speech 5

✓ Submit Rating

Results: Alignment Convergence Examples

All models achieved successful **alignment convergence**.



(a) Tacotron 2, 30 spk

(b) Tacotron 2, 60 spk

(c) Tacotron 2, 180 spk

(d) Glow-TTS, 30 spk

(e) Glow-TTS, 60 spk

(f) Glow-TTS, 180 spk

Results: Objective Evaluation

Objective metrics calculated on a 60-sentence test set (excluded from training data) by comparing synthesized audio to ground truth. Lower is better for both metrics.

Model	**Speakers**	**MCD (dB)**	F_0 **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**

Note: Minimal variation across strategies within each architecture.

Results: Subjective Naturalness (MOS)

Listening test with 21 native Lithuanian speakers (1,260 ratings). Average MOS per model architecture and data strategy shown below:

Strategy	Tacotron 2	Glow-TTS
Depth *(30 × 45.0 min)*	3.11 ± 0.16	2.13 ± 0.12
Balance *(60 × 22.5 min)*	**3.12 ± 0.17**	**2.18 ± 0.15**
Breadth *(180 × 7.5 min)*	3.03 ± 0.18	2.03 ± 0.14
Ground Truth	**4.84 ± 0.06**	

Observations

- **No significant difference** between data selection strategies within any architecture.
- Hypothesis: Rejected.
- Tacotron 2 significantly outperforms Glow-TTS in naturalness.

Results: Speaker-Dependent Naturalness

Speaker ID	Tacotron 2 MOS	Glow-TTS MOS
AS009	**4.17 ± 0.20**	**2.61 ± 0.23**
IS031	3.26 ± 0.20	2.13 ± 0.19
IS038	3.48 ± 0.21	2.51 ± 0.19
MS052	2.26 ± 0.17	1.88 ± 0.16
VP131	2.43 ± 0.19	1.93 ± 0.16
VP427	2.92 ± 0.22	1.60 ± 0.14

Observations

AS009 consistently yields highest MOS scores across all models and strategies, while *MS052*, *VP131* consistently perform poorly.

Qualitative analysis revealed noticeable muffling and reverberation in *MS052*, *VP131* recordings, pointing to data quality issues.

Conclusions

1. For Lithuanian multi-speaker TTS, within the tested range, the specific distribution of speakers does not meaningfully affect quality if the total budget is fixed.
2. 7.5 minutes per speaker is sufficient for convergence of Tacotron 2 and Glow-TTS.
3. Tacotron 2 outperforms Glow-TTS for naturalness in Lithuanian, likely due to better prosody/pitch modeling.
4. Audio quality (reverberation, muffling) and/or speaker characteristics have a strong impact on synthesis naturalness.

Recommendations & Future Work

Recommendations:

- Prioritize maximizing **total data volume** and ensuring high per-speaker quality rather than optimizing the breadth-depth balance.
- Multi-speaker TTS models can be successfully trained from scratch using sparse data with 7.5 min/speaker.

Future Work:

- Investigate more architectures (e.g., VITS, FastSpeech 2).
- Test the "breadth versus depth" trade-off at higher budgets (e.g., 100+ hours).
- Implement automated data quality filtering to remove muffled/reverberant speakers.

Thank You!

Thank you for your attention!

Appendix 1: Reviewer's Questions

- **Q1:** What is the purpose of a pre-processing step audio resampling process?
- **Q2:** Where were the model parameters obtained or how were they selected? Were they optimal, suboptimal, or random?
- **Q3:** Why the explored models (Tacotron 2 and Glow-TTS) were based on different loss function? Are these models (with different loss functions) comparable?

Appendix 2: Loss Functions

Tacotron 2 optimizes Multi-component Regression Loss:

$$\mathcal{L}_{T2} = \mathcal{L}_{Dec(L2+SSIM)} + \mathcal{L}_{Post(L2+SSIM)} + \lambda_{attn}\mathcal{L}_{Guided} + \lambda_{stop}\mathcal{L}_{Stop} \quad (1)$$

Glow-TTS optimizes Negative Log-Likelihood:

$$\mathcal{L}_{Glow} = -\sum_{j=1}^{T_{mel}} \left[\log \mathcal{N}(z_j; \mu, \sigma) + \log \left| \det \frac{\partial z_j}{\partial x_j} \right| \right] \quad (2)$$

Appendix 3: Latin Square Design Example

Listener group (ID)	Order 1	Order 2	Order 3	Order 4
Group 1	**Model A** Sentence 1	**Model B** Sentence 2	**Model C** Sentence 3	**Model D** Sentence 4
Group 2	**Model B** Sentence 3	**Model A** Sentence 4	**Model D** Sentence 1	**Model C** Sentence 2
Group 3	**Model C** Sentence 4	**Model D** Sentence 3	**Model A** Sentence 2	**Model B** Sentence 1
Group 4	**Model D** Sentence 2	**Model C** Sentence 1	**Model B** Sentence 4	**Model A** Sentence 3

Presentation order (sequence)

- 🟥 Model A
- 🟩 Model B
- 🟦 Model C
- 🟧 Model D