Enhancing Sinhala Text-to-Speech with End-to-End VITS Architecture

Sasangi Nayanathara, Inuri Harischandra, Thamira Weerakoon, Randil Pushpananda

University of Colombo School of Computing, Colombo, Sri Lanka

kksnayanathara@gmail.com, harischandra.inuri@gmail.com, thamiraweerakoon@gmail.com, rpn@ucsc.cmb.ac.lk

Introduction

Sinhala, spoken by millions in Sri Lanka, remains underexplored in **TTS** research, with existing systems producing robotic and less intelligible speech. We present the first Sinhala TTS system based on **VITS**, an end-to-end architecture with a Sinhala-specific text preprocessing pipeline and single-speaker and multi-speaker training. Experiments demonstrate state-of-the-art naturalness and intelligibility, establishing a benchmark that significantly surpasses previous Sinhala TTS approaches.

Methodology

📌 Dataset – Pathnirvana Sinhala TTS

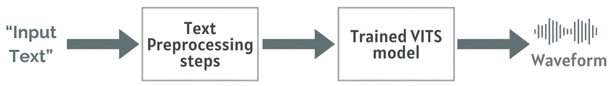
- **Sources**: Buddhist discourses & educational material
- **Format**: WAV, 22.05 kHz, 16-bit PCM
- **Splits**: Male-only, Female-only, Multi-speaker
- **Link:** https://github.com/pnfo/sinhala-tts-dataset

Table 1 - Dataset composition for Sinhala TTS training

Dataset	Speaker Name	Audio Duration	Clip Count
Male-only	Ven. Mettananda	11.8 h	5400
Female-only	Mrs. Oshadi	9h	4285
Multi-speaker	Both	20.8 h	9685

📌 Model – VITS

- **Fully End-to-end** (text → waveform) generation
- Combines text processing, acoustic modeling, and vocoder
- Robust for **low-resource languages**



Proposed TTS Process Pipeline

📌 Text Preprocessing.

Table 2 – Text Preprocessing techniques applied before synthesis

Technique	What It Handles	Example
Abbreviation Expansion	Converts Sinhala abbreviations to spoken form	පෙ.ව. → පෙරවරු *English: a.m. → ante meridiem (before noon)*
Currency Handling	Detects currency symbols (රු., $, £) and expands to spoken form	රු.150.50 → රුපියල් එකසිය පනහයි සත පනහ *English: Rs.150.50 → One hundred and fifty rupees and fifty cents*
Decimal Point Handling	Replaces numeric periods with දශම (decimal) in speech	12.5 → දොළහයි දශම පහ *English: 12.5 → Twelve point five*
Number Expansion	Breaks down and expands large numbers into natural spoken Sinhala	123456 → එකසිය විසි තුන් දහස් හාරසිය පනස් හය *English: 123456 → One hundred twenty-three thousand four hundred fifty-six*

Experiments

📌 Training Setup

- **Hardware:** 4× NVIDIA RTX 2080 Ti, 128 GB RAM
- **Batch:** 16 (train), 32 (eval); Max audio: 15 sec
- Mixed precision (fp16)
- **Logging**: Every 50 steps
- **Checkpoints**: Every 600 steps

📌 Configurations

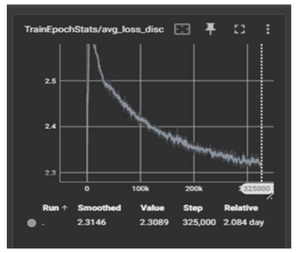
- **Single-speaker Male**
- **Single-speaker Female**
- **Multi-speaker (Male + Female)**

Table 3 - Training configurations and key observations

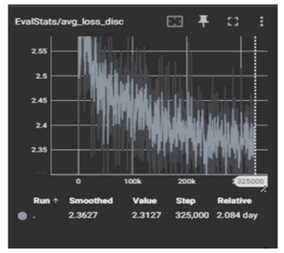
Model Configuration	Speaker (s)	Dataset	Epochs	Example
Single-Speaker Male	Male	Male-only	1,000	Stable convergence, strong intelligibility
Single-Speaker Female	Female	Female-only	4,000	Needs more data, fluctuations in loss, acceptable quality
Multi-Speaker	Male + Female	Multi-speaker	1,000	Best generalization, smooth training, supports multiple voices

📌 Monitoring

- **Training & validation loss curves**
- **Duration prediction & adversarial loss tracking**
- **All metrics monitored using TensorBoard**

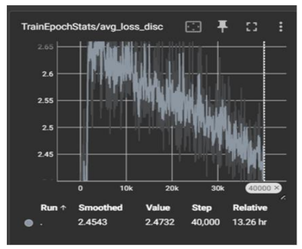


Training Loss

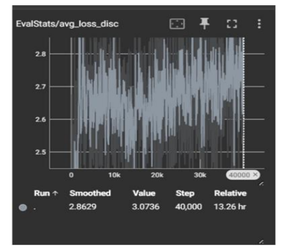


Validation Loss

Single-Speaker Male - Training and validation loss curves

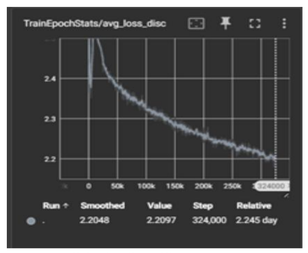


Training Loss

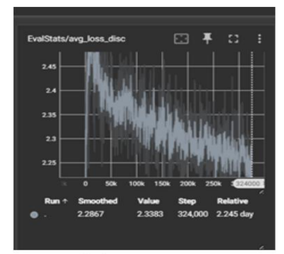


Validation Loss

Single-Speaker Female - Training and validation loss curves



Training Loss



Validation Loss

Multi-Speaker - Training and validation loss curves

Results

📌 Subjective Evaluation

- **Mean Opinion Score (MOS)**
  - ✓ 5-point Likert scale,
  - ✓ 12 native speakers (6M, 6F),
  - ✓ 15 Sinhala sentences (short/medium/long).
- **Semantically Unpredictable Sentences (SUS)**
  - ✓ 10 sentences per listener
  - ✓ word accuracy used to measure intelligibility.

📌 Objective Evaluation

- **Mel Cepstral Distortion (MCD)**
  - ✓ Computed between generated and reference audio features
  - ✓ Measures spectral similarity → lower = better

Table 4 – Subjective & Objective Evaluation Results

Model Configuration	MOS (%)		SUS(%)	MCD (dB)
	Intelligibility	Naturalness		
Single-Speaker Male	**92.33**	**83.54**	**85.83**	**13.27**
Single-Speaker Female	71.71	74.50	77.50	20.56
Multi-Speaker Male	81.33	79.56	82.50	14.07
Multi-Speaker Female	**84.78**	**81.44**	**81.39**	**20.29**

Single-speaker male achieved the best, intelligibility, while multi-speaker models improved generalization and female voice quality.

Table 5 - Comparison with Prior Sinhala TTS

Sinhala TTS System	MOS (%) Naturalness	SUS (%) Intelligibility
Nanayakkara et al. (2018)	70	70
Tacosi (2023)	84.00	78.2
Sinhala VITS (Proposed)	**85.83**	**82.34**

Conclusion

- First **Sinhala VITS-based TTS system**.
- Achieved state-of-the-art results:
  - Intelligibility – 85.8% (SUS)
  - Naturalness – 83.54% (MOS)
- Outperforms prior Sinhala TTS approaches.
- Establishes a benchmark for future research.
- Future work: multilingual pretraining, speaker adaptation, dataset expansion.

Acknowledgement

Thanks to the UCSC Language Technology Lab, especially Dr. A.R. Weerasinghe, for their valuable guidance.