Survey on Neural Text-to-Speech Systems for Sinhala Language

Visitha Wickramasinghe, Kusalya De Zoysa, Haritha Mihimal, Fathima Hana

Abstract

Sinhala, the official language of Sri Lanka, has long faced challenges in developing high-quality text-to-speech (TTS) systems due to limited datasets, linguistic complexity, and a lack of open resources. Traditional rule-based or concatenative approaches like achieved only modest intelligibility. With the emergence of neural architectures Sinhala TTS has seen significant improvement in fluency and naturalness. However, there are still critical gaps: no Sinhala TTS system today supports real-time streaming, voice cloning, or speaker adaptation. This survey reviews the evolution from classical methods to recent neural models for Sinhala, highlighting strengths, limitations, datasets, and research gaps to support future work in the field.

1 Introduction

Sinhala is the mother tongue of approximately 74% of Sri Lanka's 22 million people. High quality Sinhala TTS is essential for accessibility (e.g., screen readers), education, and media localization. Early Sinhala TTS systems were rule-based or concatenative and suffered from limited naturalness, robotic prosody, and narrow vocabulary. The last decade has brought neural TTS architectures that outperform traditional methods and require fewer handcrafted rules. This survey summarizes current Sinhala TTS research, focusing on neural approaches, datasets, evaluations, and remaining challenges.

2 Sinhala Language and Classical TTS

Sinhala's writing system is a Brahmi-derived script with complex syllabic and phonetic rules. A particular challenge is the pervasive schwa epenthesis: consonants often require an implicit vowel sound [a] to be inserted or omitted, which must be captured by grapheme-to-phoneme (G2P) rules [1] . Early work addressed these front-end issues: for example, a Sinhala G2P conversion with explicit schwa rules achieved 98% accuracy on a 30k-word test set laying groundwork for TTS. Traditional concatenative systems used large diphone or unit databases [1].

The earliest Sinhala TTS effort was developed by Weerasinghe et al. in 2007, who introduced Festival-SI, a unit-selection concatenative system built on the Festival framework. It employed a method where recorded speech was segmented into phonetic units and recombined during synthesis. While the system achieved real-time playback on standard hardware, its output was rated with a MOS (Mean Opinion Score) of approximately 3.3, indicating intelligibility but lacking naturalness, fluency, and emotional tone [2].

Nanayakkara et al. extended the MaryTTS platform to support Sinhala by recording a custom Sinhala dataset and developing a grapheme-to-phoneme converter specifically for the language [1]. This resulted in a higher intelligibility and naturalness compared to Festival-SI. Their contribution demonstrated that open-source platforms could be adapted to support underrepresented languages like Sinhala with minimal computational resources.

Subsequently, Lakmal et al. and Senarathna et al.worked on further enhancing the MaryTTS Sinhala voices by defining text normalization rules, phoneme mappings, and prosody modeling using HMM-

based synthesis. These voices were implemented in lightweight systems intended for accessibility tools. However, their quality remained limited by monotonic intonation, a lack of speaker diversity, and no fine-tuning mechanisms for new speakers or emotions [3][4].

Although these concatenative systems made Sinhala TTS possible, they suffered from limited naturalness and prosody. Weerasinghe et al. note that even with a 10-million-word corpus, early Festival/Mary TTS voices lacked natural intonation and gender diversity [1]. In practice, spoken Sinhala TTS remained "robotic" and struggled with context like numbers, abbreviations, or mixed-language (English/Sinhala) text. Practical challenges (scarce corpora, regional dialects, code-switching) have been highlighted [5].

3 Neural TTS Architectures

Modern neural TTS models generate speech end-to-end from text via deep learning. Key architectures include:

Tacotron (2017) is a sequence-to-sequence model with attention that maps characters (or phonemes) to mel-spectrogram frames. Tacotron (and its successor Tacotron 2) require no hand-crafted features and learn alignments automatically. Output spectrograms are converted to waveform by a vocoder (e.g. Griffin-Lim in Tacotron, or neural vocoders like WaveNet/Vocoder) [5].

WaveNet (2016) is a powerful autoregressive neural vocoder by van den Oord et al. WaveNet models raw audio and can produce very natural speech when conditioned on acoustic features [6].

More recent advances like VITS (variational inference with adversarial training) and FastSpeech/FastSpeech2 (non-autoregressive transformer TTS) further accelerate training and inference. These use phoneme encoders, duration predictors, and parallel generation for faster synthesis.

In general, neural TTS systems outperform concatenative methods in naturalness and allow easier multi-speaker support. However, they typically require large, high-quality datasets. Below we review the main Sinhala TTS efforts using neural models.

4 Neural TTS Systems for Sinhala

4.1 TacoSi (2023) – Tacotron-Based Sinhala TTS

TacoSi (Kasthuriarachchi et al.) is a neural Sinhala TTS based on the Tacotron 2 architecture. It is trained end-to-end on raw Sinhala text–speech pairs. A sequence-to-sequence Tacotron model produces mel-spectrograms from graphemes (Sinhala Unicode). Griffin–Lim vocoder is used to invert to waveform. The authors collected a Sinhala corpus (size not explicitly stated, but presumably several hours). They emphasize raw text training without handcrafted features. In listening tests (10 native speakers, 10 sample utterances), TacoSi achieved a Mean Opinion Score (MOS) of 4.39 (out of 5), significantly higher than prior systems. Intelligibility (comprehensibility) reached 84%, "significantly higher" than the 70% reported for older systems. Importantly, TacoSi handled rare words and numeric expressions well, a key advantage of end-to-end models [5].

These results indicate that neural TTS can produce human-like Sinhala speech far exceeding previous concatenative systems. TacoSi's 4.39 MOS suggests listeners judged the speech nearly on par with natural recordings. The authors attribute improvements to not using phoneme alignments and relying on data-driven training [5].

4.2 Deep-Voice TTS (2019) – Fully Convolutional Approach

Jayawardhana et al. implemented a Sinhala-English TTS using a Deep Voice style neural model. This system is an attention-based, fully-convolutional neural TTS (inspired by Deep Voice 1/2) that synthesizes Sinhala speech [7]. The model follows Deep Voice– an encoder-decoder with convolutional layers and attention. It does not rely on LSTMs. The authors noted that a very large dataset is needed for

intelligible synthesis. They estimate at least 1,000 utterances (≈96 hours) of Sinhala speech are required for good quality. (For comparison, they report English requires 1,000 utterances ≈24 hours). No explicit MOS was reported, but the implication is that building a competitive system is data-intensive. The 2019 system likely served as a proof-of-concept.

This work highlights a common issue: neural TTS needs large amounts of data. In practice, 96 hours for 1,000 utterances is unusually high (1000 × 10s ≈2.8h, so perhaps "96h" is a modeling of training duration). It suggests that with limited Sinhala data, achieving natural synthesis is hard. (The TacoSi result implies that more efficient models can perform well with less data [7].

4.3 Path Nirvana Sinhala TTS (2023) – VITS-Based Neural Model

In 2023, the Path Nirvana Foundation released a publicly accessible Sinhala TTS model built using the VITS (Variational Inference Text-to-Speech) architecture. VITS is a modern, end-to-end neural TTS model that combines text encoding, speech decoding, and waveform generation into a unified system. Unlike Tacotron-based pipelines that rely on external vocoders (e.g., Griffin-Lim), VITS includes a built-in neural vocoder, enabling both faster and higher-quality synthesis [11].

Based on VITS, which uses variational autoencoding and adversarial training for realistic speech generation. It learns duration, pitch, and text-to-speech alignment implicitly. Trained on 11.8 hours of high-quality Sinhala speech, recorded by a single male speaker. The dataset, also released by the Path Nirvana team, includes phonetically balanced sentences designed to cover syllabic diversity [11].

The voice quality is Subjectively rated to be highly natural, though formal MOS evaluations were not published at the time of release.Being a single-speaker model, it lacks native support for speaker cloning or emotional expressiveness. However, the voice is consistent, smooth, and usable in production .

5 Datasets and Resources

High-quality speech data is critical. Fortunately, Sinhala has seen new resources. Google SLR30 (SLR 30): A public dataset of multi-speaker transcribed Sinhala speech collected by Google. It contains 699 MB of WAV files and corresponding transcripts. Since it is CC BY-SA 4.0 licensed, it is a key resource for training neural models [8].

Path Nirvana Sinhala TTS Dataset: The Path Nirvana Foundation (a Sri Lankan nonprofit) released an open Sinhala TTS corpus in 2019 (3300 sentences, 7.5h) and an expanded version in 2023 (6248 utterances, 13.8h). This multi-speaker dataset (male/female) was recorded to cover rare syllables in Sinhala. It is freely available on GitHub [9].

Other speech corpora: Additional Sinhala speech data exists (e.g. OpenSLR ASR sets, commoncrawl reads) but not specifically TTS-aligned. In domain-specific work, a crowdsourced "Voicer" corpus for banking phrases (10h) has been collected. However, the above two (SLR30 and Path Nirvana) are the primary open datasets used for TTS [10].

Lexicons and Grapheme Rules: Open pronunciation dictionaries and Sinhala phonology descriptions also aid TTS. Wasala et al. provide Sinhala grapheme-to-phoneme rules. A full lexicon of Sinhala pronunciations (for MARY TTS) is maintained by the University of Colombo group, but it is not freely published [12].

These resources have enabled the recent neural models. For example, TacoSi likely drew on published corpora, and Path Nirvana's dataset is specifically intended for such models. The combination of data availability and open frameworks (TensorFlow/PyTorch TTS libraries) now makes development of Sinhala TTS more feasible than ever.

6 Evaluation and Comparison

The reported evaluations of Sinhala TTS systems show dramatic improvements with neural methods (Table 1). Older systems like Festival-si or MARY TTS achieved intelligibility around 70%. In contrast, TacoSi (Tacotron) reaches 84% intelligibility and human-level MOS. This is summarized below with their strengths and weaknesses.

System	**Year**	**Approach**	**Data**	**Intelligibility / MOS**	**Strengths**	**Weaknesses**
Festival-SI	2007	Concatenative	10k word corpus	71.5% (MRT)	First Sinhala TTS; real-time capable	Robotic voice; no flexibility
MaryTTS Sinhala	2018	Unit Selection	1000 sentences	70%	Screen reader support, higher quality than Festival	Fixed voice, lacks naturalness and emotion
Deep Voice	2019	Conv Seq2Seq	1000 utterances	Not reported	Early neural baseline	Very high data need
TacoSi	2023	Tacotron2	Several hours	84% intelligibility; MOS 4.39	Handles rare cases; strong prosody	Extremely high latency; no real-time
Path Nirvana VITS	2023	VITS	11.8h single-speaker	Not reported	High quality output	No cloning; single voice; lacks emotion

Table 1: Sinhala TTS systems and results. Intelligibility is typically measured by listener transcription tests (Modified Rhyme Test or SUS), while MOS is the mean opinion score for naturalness.

7 Open Challenges and Future Directions

Despite significant progress in neural Sinhala TTS, several open challenges remain. Although new datasets have emerged, high-quality, speaker-consistent recordings are still limited. Most systems rely on single-speaker corpora. Neural models typically require 10+ hours per speaker, and the lack of diversity across age, gender, and dialects restricts generalizability. While the Path Nirvana 13.8h dataset is a major step forward, more multi-speaker, child, and dialectal data is needed [9].

Sinhala's phonological structure, including schwa epenthesis, consonant clusters, and vowel insertion rules, presents difficulties for end-to-end models. While some neural models learn these patterns implicitly, explicit G2P components could further improve output quality.

There is no standard Sinhala TTS benchmark. Reported MOS scores are not always comparable due to variations in evaluation methodology. A shared Sinhala SUS or MRT sentence set and standardized MOS protocols would benefit future comparisons.

No Sinhala TTS system has demonstrated integrated real-time ASR $\rightarrow$ MT $\rightarrow$ TTS streaming pipelines suitable for live dubbing, assistive devices, or conversational agents.

All current Sinhala TTS systems are single-voice and lack speaker identity preservation. There is no open-source support for cloning a user's voice, adapting models to new speakers, or enabling speaker consistency across translations — a vital requirement for dubbing and personalization.

To fully unlock the potential of Sinhala TTS, future work must go beyond single-speaker offline synthesis and embrace more expressive, real-time, and adaptive technologies. First, researchers should explore advanced architectures such as FastSpeech2, VITS, or Glow-TTS with style tokens, which allow parallel

generation and greater prosodic variation. These models can improve both the speed and naturalness of Sinhala speech output, making them well suited for real-time dubbing or interactive systems.

Second, multilingual transfer learning represents a promising path forward. Given the linguistic similarities between Sinhala and related Indo-Aryan languages like Hindi, models pre-trained on larger corpora in these languages could be fine-tuned on smaller Sinhala datasets, accelerating development and improving quality in low-resource conditions.

Third, future systems should prioritize modular and real-time pipelines that integrate TTS with Automatic Speech Recognition (ASR) and Machine Translation (MT) components. This integration is essential for applications like live dubbing, real-time language translation, and conversational AI, which are currently unavailable for Sinhala.

Finally, a major research frontier lies in voice personalization. By supporting techniques like speaker embeddings, voice cloning, and zero-shot adaptation (as demonstrated by Seed-VC), Sinhala TTS can evolve from single-speaker systems to dynamic, user-tailored solutions that maintain speaker identity across languages and contexts.

The inclusion of real-time streaming synthesis and voice preservation capabilities would represent a transformative step in Sinhala TTS. It would enable not only better human-computer interaction but also expand the use of Sinhala in global accessibility technologies, education platforms, and multimedia content localization—domains where the Sinhala language has so far remained underserved.

References

[1] L. Nanayakkara, C. Liyanage, P. Viswakula, T. Nadungodage, R. Pushpananda, R. Weerasinghe, *Development of a Sinhala Voice for the MaryTTS Text-to-Speech Platform*, Proceedings of MERCon, 2018.

[2] Weerasinghe, R. T. Kasthuriarachchi, *TacoSi: Sinhala TTS using Tacotron 2*, Proceedings of MERCon, 2023.

[3] D. Lakmal et al., *Rule-based Sinhala Grapheme-to-Phoneme Converter for MaryTTS*, 2019.

[4] P. Senarathna et al., *Sinhala Text Normalization and Prosody Modeling for HMM-based Synthesis*, 2019.

[5] T. Kasthuriarachchi, *TacoSi: A Sinhala Text to Speech System with Neural Networks*, 2023.

[6] A. van den Oord et al., *WaveNet: A Generative Model for Raw Audio*, 2016.

[7] P. Jayawardhana, A. Aponso,N. Krishnarajah, A. Rathnayake, *An Intelligent Approach of Text-To-Speech Synthesizers for English and Sinhala Languages*, 2019.

[8] OpenSLR, `https://openslr.org/30/`

[9] Path Nirvana Sinhala Dataset, `https://github.com/pnfo/sinhala-tts-dataset`

[10] De Silva Nisansa, *Survey on Publicly Available Sinhala Natural Language Processing Tools and Research*, 2024.

[11] Path Nirvana GitHub, `https://github.com/pathnirvana/coqui-tts`

[12] Asanka Wasala, Ruvan Weerasinghe, and Kumudu Gamage. 2006. *Sinhala Grapheme-to-Phoneme Conversion and Rules for Schwa Epenthesis*, In Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 890–897, Sydney, Australia. Association for Computational Linguistics.