Project Proposal: Real-Time English-to-Sinhala Dubbing System

Supervised by: Dr. Uthayasanker Thayasivam
Department of Computer Science and Engineering
University of Moratuwa

June 7, 2025

Abstract

The growing presence of English audiovisual content poses a language barrier for Sinhala-speaking audiences in Sri Lanka. While some commercial dubbing solutions exist, they often lack real-time capabilities, voice personalization, or developer accessibility. This project introduces a Real-Time English-to-Sinhala Dubbing System featuring both offline and real-time pipelines. It integrates speech recognition, machine translation, and a dual text-to-speech (TTS) approach with voice preservation capabilities. The system also supports background audio separation, speaker consistency, and emotional tone rendering. Designed for both uploaded and live content, this solution aims to enhance media accessibility and contribute to media accessibility, educational support and speech technology research in low-resource language contexts.

1. Introduction

1.1 Background

With the rapid globalization of media, English-language videos dominate online platforms, educational materials, and entertainment content. However, this creates an accessibility gap for Sinhala-speaking populations in Sri Lanka who may struggle with understanding English audio. While subtitles offer a partial solution, they are not ideal for younger audiences, individuals with reading difficulties, or situations where reading subtitles isn't feasible. As a result, voice-based translation, especially dubbing, has become a compelling avenue for making content inclusive and accessible. Advances in AI

have made speech recognition, translation, and speech synthesis possible, yet Sinhala remains a low-resource language with limited support for high-quality dubbing.

1.2 Motivation

Although some commercial platforms offer English-to-Sinhala dubbing services, they are often expensive, lack real-time interaction, or do not preserve speaker identity and tone. Furthermore, existing systems typically do not allow customization or access to developers for experimentation or research. Our motivation is to build an open, extensible, and research-oriented dubbing system that supports both offline and real-time scenarios, while preserving speaker consistency and emotional nuance.

1.3 Problem Statement

There is currently no publicly available system that can deliver natural-sounding, real-time Sinhala dubbing for English audiovisual content, especially with features such as voice consistency, emotion preservation, and low-latency streaming. Most existing solutions are limited to offline use, generic synthetic voices, or lack synchronization with the source content.

1.4 Research Objectives

This project aims to:

1.4.1 Develop a two-phase system:

- Phase 1: Offline dubbing of uploaded videos.
- Phase 2: Real-time dubbing for streaming or live content.

1.4.2 Implement a modular dubbing pipeline combining ASR, NMT, and TTS.

- Automatic Speech Recognition (ASR),
- Neural Machine Translation (NMT),
- Dual-mode Text-to-Speech (TTS) synthesis.

1.4.3 Provide a Sinhala TTS component with:

- Natural-sounding speech synthesis,
- Voice cloning for speaker identity preservation,
- Low-latency output suitable for offline use,
- Real-time streaming capability (Filling a major gap in current open and commercial systems, which lack these features for Sinhala)

1.4.4 Optimize for performance:

- Low latency,
- High naturalness,
- Intelligibility of Sinhala output.

1.4.5 Design a scalable and extensible architecture.

- Academic research,
- Developer experimentation,
- Media and accessibility applications.

1.5 Preliminary Suppositions and Implications

We assume that with careful integration of modern ASR, MT, and TTS technologies, it is feasible to build a Sinhala dubbing system that performs well in real-time and offline modes. We also expect that preserving speaker identity and emotion will significantly enhance realism and user engagement. If successful, this project will contribute a novel solution for Sinhala media localization and provide a foundation for further research in speech technology for low-resource languages.

2. Related Works

The earliest Sinhala TTS effort was developed by Weerasinghe et al. (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 [1].

Later, Nanayakkara et al. (2018) extended the MaryTTS platform to support Sinhala by recording a custom Sinhala dataset and developing a grapheme-to-phoneme converter specifically for the language. 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 [2].
Subsequently, Lakmal et al. and Senarathna et al. (2019) 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].

A major leap in quality came with the introduction of TacoSi by Weerasinghe et al. (2023), a Sinhala neural TTS system based on the Tacotron 2 architecture. TacoSi employed a sequence-to-sequence model to convert input text to mel-spectrograms, followed by a Griffin-Lim vocoder for waveform synthesis. It achieved an MOS of 4.39 and an 84% intelligibility rate in native speaker evaluations, far surpassing earlier systems. However, the Griffin-Lim vocoder used in the pipeline caused synthesis to be slow—around 30 seconds for a 6-second sentence—making it unsuitable for real-time applications without additional optimization also it does not provide any voice cloning capabilities. [5].

The Path Nirvana team released a VITS-based Sinhala TTS model in 2023. Unlike TacoSi, this model integrated the vocoder into the synthesis architecture. It was deployed using runtime-optimized backends like ONNX and Piper for fast inference. Despite these advantages, the model was single-speaker only, and lacked mechanisms for voice cloning, speaker adaptation, or expressive synthesis, such as emotional tone variation [6].

The first MT system involving Sinhala was SEES (Sinhala to English and English to Sinhala), introduced by Wijerathna et al. (2012). It was a rule-based system that leveraged a bilingual lexicon and manually crafted grammatical rules to perform two-way translation. Impressively, it also supported transliterated Sinhala written in Latin script ("Singlish"). SEES served as a valuable proof of concept but suffered from rigid syntax handling, inability to generalize to complex inputs, and poor fluency due to its symbolic nature [7].

A more scalable approach was introduced by Perera et al. (2022), who experimented with incorporating linguistic features—specifically part-of-speech (POS) tags—into Transformer-based Neural Machine Translation (NMT) models for English to Sinhala translation. They found that augmenting the source input with syntactic information led to significantly improved grammatical correctness and lexical choices in the Sinhala output, addressing issues of syntactic divergence and morphological complexity in this low-resource language pair [8].

In terms of commercial tools and platforms, Wavel AI is a commercial dubbing platform offering English-to-Sinhala translation through automated pipelines. It allows users to upload video content and select from predefined Sinhala voices. It operates only in offline mode and does not provide real-time synthesis, developer customization, or open-source integration [11].

VideoDubber is another tool that offers Sinhala voice dubbing. It promotes speaker tone preservation and automatic synchronization with source speech, including limited voice cloning capabilities [12]

Kapwing and Checksub are content editing tools that incorporate basic TTS dubbing for multiple languages, including Sinhala. These tools rely on cloud TTS APIs (e.g., Microsoft Azure) to generate audio, typically using one or two predefined voices per language. Customization, emotional tone control, and real-time streaming are not supported, and no internal TTS models are available to users [13].

In terms of existing dubbing extensions, LingoCub is a Chrome extension that enables real-time dubbing of YouTube videos using ElevenLabs APIs. It performs speech recognition, translation, and TTS synthesis on the fly, synchronizing output with the video timeline. However, ElevenLabs does not support Sinhala, making the tool unusable for Sinhala content despite its strong real-time capability for other languages [14].

Transmonkey integrates Whisper for ASR, LLMs for translation, and OpenAI TTS for speech synthesis. It supports real-time dubbing across many languages with customizable tone and style. Nevertheless, Sinhala is not available due to the lack of compatible, natural-sounding Sinhala TTS voices in OpenAI's current offerings [15].

YouTube Dubbing Translate & Dub is a browser extension that overlays AI-generated

voice translations on YouTube content. While it supports subtitle-based translation in Sinhala, its TTS voice output for Sinhala is limited to robotic voices or fails entirely, as it depends on browser-native or Azure voices without emotional nuance or voice cloning [16].

Considering other multilingual TTS platforms, OpenAI released their newest API in March 2025 for TTS purposes, called GPT-4o-mini-TTS. As part of OpenAI's voice synthesis suite, it can generate highly realistic speech across many languages. However, while Sinhala text input is supported, the output is restricted to a few fixed predefined voices[17].

Google released Gemini 2.5 Flash Preview in May 2025 which also introduces voice interactivity and cross-language comprehension. However, voice output in Sinhala remains unavailable or experimental. Like GPT-4o-mini-TTS, it lacks voice personalization, or cloning [18].

Microsoft Azure TTS does provide high-quality Sinhala voices such as "Thilini" and "Sameera," with expressive modeling. However, these voices are static and cannot be fine-tuned or cloned. Developers cannot modify or retrain them for specific domains or speaker identities. Thus, while suitable for generic TTS tasks, they do not meet the personalization and modularity needs for advanced dubbing systems [19].

Although research and commercial interest in Sinhala dubbing is steadily growing, no current system offers a comprehensive solution that meets all key requirements for high-quality dubbing in Sinhala. Specifically, the following capabilities are missing from all known systems:

- Real time dubbing - No tool supports real-time English-to-Sinhala dubbing with natural voice output,
- Speaker identity preservation - Most Sinhala TTS systems use a fixed voice, none clone or adapt the speaker's identity
- Low latency, natural TTS with fine-tuning - While Path Nirvana VITS offers low latency, it lacks speaker adaptation. TacoSi is natural, but too slow. Commercial APIs provide speed but no control.
- Open source modularity and developer control - There is no complete, modular, open-source Sinhala dubbing pipeline combining ASR, MT, and TTS with real-time capability and tunability.

3. Proposed Approach

To build a high-quality English-to-Sinhala dubbing system that supports real-time and offline processing, we adopt a modular pipeline design. This architecture integrates Automatic Speech Recognition (ASR), Neural Machine Translation (NMT), and Text-to-Speech Synthesis (TTS) with speaker-aware and emotion-aware extensions. One of the core components of this project is the development of a fine-tuned Sinhala TTS model using XTTS_v2.

3.1 Fine-Tuning XTTS_v2 for Sinhala TTS

One of the primary goals of this project is to build a real-time-capable Sinhala text-to-speech (TTS) model by fine-tuning XTTS_v2, a multilingual and speaker-adaptive neural TTS architecture. While XTTS_v2 supports multilingual synthesis and voice cloning, Sinhala is not natively supported. To enable high-quality Sinhala speech generation, we fine-tuned the model using the 11.8-hour single-speaker Sinhala dataset publicly released by Path Nirvana in 2023.

This dataset contains studio-quality recordings of a male speaker, covering a wide range of phonemes, sentence structures, and prosodic patterns in Sinhala. The dataset was preprocessed with silence trimming, normalization, and forced alignment to ensure compatibility with XTTS_v2's training pipeline.

Key aspects of the fine-tuning process include:

- Adaptation of XTTS_v2's multilingual layers to accept Sinhala phoneme sequences and align them with mel-spectrogram outputs.
- Training conducted on GPU infrastructure (e.g., NVIDIA T4) with memory-efficient settings and checkpointing.
- Evaluation using metrics such as MOS (Mean Opinion Score), intelligibility, and inference speed, particularly with real-time inference backends like ONNX.

Although the dataset is single-speaker only, the fine-tuned model delivers highly natural Sinhala speech, suitable for default dubbing output. For voice cloning, the system integrates an external module (e.g., Seed-VC) to transfer target speaker identity post-synthesis.

This customized XTTS_v2 serves as the default TTS engine in both Phase 1 (offline) and Phase 2 (real-time) pipelines, offering low-latency, intelligible, and expressive Sinhala output.

3.2 Phase 1 – Offline Dubbing System

The offline pipeline is designed to process complete audio or video files and generate Sinhala-dubbed output while preserving timing, speaker characteristics, and audio quality. It includes the following components:

- **Audio Separation:** Isolates clean speech from background music or effects for accurate transcription and reintegration.
- **ASR – Automatic Speech Recognition:** English speech is transcribed using FastWhisper with Silero VAD for precise segmentation.
- **MT – Neural Machine Translation:** English to Sinhala translation using Meta NLLB-200 (1.3B) and CTranslate2.
- **TTS – Text-to-Speech Synthesis (Dual Approach):**
  - Option 1: Fine-tuned XTTS_v2 model.
  - Option 2: GPT-4o-mini-TTS + Seed-VC for expressive, cloned output.
- **Audio Postprocessing:** Time-stretching to synchronize Sinhala speech with original audio.

This pipeline enables accurate and natural Sinhala dubbing of full-length videos with attention to timing, speaker identity, and overall quality.

3.3 Phase 2 – Real-Time Streaming Dubbing System

The real-time system extends the Phase 1 pipeline to work on live or streaming inputs with low latency. It is designed as a pipelined multi-stage architecture consisting of three asynchronous and independent modules—ASR, MT, and TTS—each processing different chunks concurrently.

3.3.1 Streaming Pipeline and Chunking Logic

- **Sentence-Level Chunking:** Input audio is divided into sentence-level chunks (not fixed time windows), using Silero VAD and FastWhisper to ensure natural linguistic boundaries. Each chunk contains a single speaker, enabling consistent voice identity and simplifying TTS voice assignment.
- **Pipelined Processing:**
  - Chunk 1 is being processed by the TTS stage

  - Chunk 2 is simultaneously translated by the MT stage,
  - Chunk 3 is transcribed by the ASR stage.(Each stage is independent, allowing parallelism and reducing overall end-to-end latency.)

- **Model Flow:**
  - ASR: FastWhisper + Silero VAD,
  - MT: Meta NLLB-200 1.3B + CTranslate2,
  - TTS: XTTS_v2 or GPT-4o-mini-TTS + Seed-VC.

- **Inter-Stage Buffers:** To handle speed variations between modules, buffer queues are added between stages. This allows asynchronous data flow without bottlenecks or drops, ensuring smooth chunk handoff and continuous dubbing.

3.3.2 Time Stretching and Audio Sync

- The synthesized Sinhala audio chunk may be slightly longer or shorter than the original English chunk due to structural differences in translation. To address this, the final output is dynamically stretched or compressed to match the timing of the source, using audio time-stretching algorithms. This guarantees alignment with original video frames or background audio.

- A delay buffer (e.g., 2–3 seconds) is introduced to allow the pipeline to prepare dubbed audio slightly ahead of playback, ensuring near-real-time dubbing with minimal desynchronization.

3.4 Modularity and Extensibility

Each module ASR, MT, TTS is encapsulated and independently operable. This enables:

- Swapping or fine-tuning components without altering the rest of the system.

- Deployment as microservices or streaming agents.

- Integration with other applications like browser extensions, video players, or assistive devices.

The system is optimized for GPUs (e.g., NVIDIA T4), supports quantized inference, and is scalable for other languages or voice settings.

4. Expected Outcomes and Conclusion

By the end of this project, we expect to deliver the following key outcomes:

4.1 Fine-Tuned Sinhala XTTS_v2 Model

- A real-time capable Sinhala TTS model fine-tuned from XTTS_v2 using the 11.8-hour single-speaker dataset released by Path Nirvana.
- The model will serve as the system's default synthesis engine and support high-quality, natural Sinhala speech generation with voice preservation suitable for dubbing.
- The first open source Sinhala TTS model with real time streaming and voice cloning functionalities.

4.2 Phase 1: Functional Offline English-to-Sinhala Dubbing System

- A complete offline pipeline capable of converting pre-recorded English audio or video into natural, intelligible Sinhala speech.
- Integration of background audio separation and re-merging for professional-quality dubbing.
- Support for speaker-aware synthesis using voice cloning and optional emotion modeling.
- A dual-path TTS module offering high-quality synthesis via a fine-tuned XTTS_v2 model and an alternative GPT-4o-mini-TTS + Seed-VC route.

4.3 Phase 2: Real-Time Streaming Dubbing System

- A modular, low-latency streaming system that supports live English audio input and near-real-time Sinhala output.
- Pipelined ASR $\rightarrow$ MT $\rightarrow$ TTS architecture with sentence-level chunking and inter-stage buffers.
- Dynamic audio alignment using time-stretching for seamless dubbing.
- Ability to preserve speaker identity and apply emotional tone modulation in real time (optional but planned).

4.4 Reusable and Extensible System

- A modular codebase supporting future expansion into other language pairs or voice styles.
- Components deployable as microservices or API endpoints for integration into third-party applications (e.g., Chrome extensions, accessibility platforms, media editing tools).

The increasing consumption of English audiovisual content in Sri Lanka presents a significant accessibility barrier for Sinhala-speaking audiences. While commercial dubbing tools exist, they often lack real-time performance, personalization, or openness for local innovation. This project bridges that gap by delivering a novel, modular dubbing system tailored for Sinhala, a language underrepresented in speech technology research. Through its two-phase architecture, speaker-aware synthesis, real-time processing, and careful system design, the project demonstrates how cutting-edge language models can be effectively adapted for low-resource, real-world scenarios. Beyond media accessibility, the system has broader applications in education, entertainment, and assistive technology and lays the foundation for future research in low-latency speech-to-speech translation, multilingual voice cloning, and emotion-aware dubbing.

References

1. Weerasinghe, R., et al. (2007). *FestivalSI: A Sinhala Text-to-Speech System.* Language Technology Research Laboratory, UCSC.
2. Nanayakkara, D., et al. (2018). *Development of a Sinhala Voice for the MaryTTS Text-to-Speech Platform.* Proceedings of MERCon.
3. Lakmal, D. et al. (2019). *Rule-based Sinhala Grapheme-to-Phoneme Converter for MaryTTS.*
4. Senarathna, P. et al. (2019). *Sinhala Text Normalization and Prosody Modeling for HMM-based Synthesis.*
5. Weerasinghe, R. et al. (2023). *TacoSi: Sinhala TTS using Tacotron 2.* In Proceedings of MERCon.
6. Path Nirvana (2023). *Sinhala VITS Neural TTS Model Release* [GitHub/Technical Report].

7. Wijerathna, H. et al. (2012). *SEES: Sinhala to English and English to Sinhala Rule-based Translator*. ICTer Conference.

8. Perera, P. et al. (2022). *Enhancing English-to-Sinhala NMT with POS-tag Conditioning*. ICTer Journal.

9. Shreya G. S. et al. (2022). *Speech-to-Speech Language Translator: English–Sinhala for Healthcare*. IJARIIE.

10. SLIIT NLP Group (2023). *Sinhala-English Voice Translator Project*.

11. Wavel AI. (2023). `https://wavel.ai`

12. VideoDubber. (2023). `https://videodubber.com`

13. Kapwing TTS API. (2023). `https://www.kapwing.com`

14. LingoCub Chrome Extension. (2024). `https://chrome.google.com/webstore`

15. Transmonkey. (2024). `https://www.transmonkey.ai`

16. YouTube Dubbing – Translate & Dub Extension. (2024). `https://dub.video`

17. OpenAI. (2024). *GPT-4o Voice Capabilities*. `https://openai.com/gpt-4o`

18. Google DeepMind. (2024). *Gemini 2.5 Flash Technical Preview*. `https://deepmind.google/technologies/gemini`

19. Microsoft Azure. (2024). *Text to Speech – Neural Voices: Sinhala*. `https://learn.microsoft.com/en-us/azure/ai-services/speech-service/language-support#text-to-speech`