A Real-Time English-to-Sinhala Dubbing System with Voice Preservation and Low-Latency Synthesis

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

***Abstract*—Speech to speech translation systems have gained prominence for cross-lingual media accessibility and communication. However, Sinhala—a morphologically rich, low-resource language spoken by over 17 million people in Sri Lanka—remains underrepresented in such technologies. This paper presents a modular real-time English-to-Sinhala dubbing system designed to overcome these limitations using open-source tools. Our system integrates Automatic Speech Recognition, Machine Translation and Text to Speech as a cascaded pipeline. We implement two pipelines: an offline high-quality dubbing pipeline and a real-time streaming variant with pipelined stages. The system is expected to provide intelligible, natural, and speaker-consistent Sinhala voiceovers with acceptable latency for real-time applications. This work is one of the first to provide real-time dubbing with voice preservation for Sinhala and contributes toward low-resource language accessibility and speech translation research.**

I. Introduction

In the age of global digital media, the demand for real-time cross-lingual dubbing systems has increased dramatically, particularly in the domains of education, accessibility, and entertainment. Tools such as Google Translate and ElevenLabs offer real-time text and speech services across many languages—but Sinhala, spoken by over 74

Current Sinhala dubbing methods rely primarily on manual voiceover or batch-processed synthetic speech. Prior efforts in Sinhala text-to-speech (TTS) have produced intelligible but limited systems such as Festival-SI [2] and MaryTTS Sinhala [3]. More recent neural models like TacoSi [4] and Path Nirvana [5] show improved fluency and MOS scores but are not suitable for real-time or speaker-personalized synthesis.

In the broader machine translation (MT) space, tools like Meta's NLLB-200 [6] provide high-quality translations to Sinhala, and OpenAI's Whisper [7] enables accurate English ASR. However, no existing system brings together these components into a real-time Sinhala dubbing pipeline with low latency, high intelligibility, and speaker identity preservation.

This paper addresses that gap. We propose a real-time English-to-Sinhala dubbing system built entirely from open tools and datasets. It consists of three primary stages—ASR, MT, and TTS—with real-time pipelining and optional voice cloning. The TTS stage is powered by a fine-tuned XTTS_v2 model, trained on high-quality Sinhala data, enabling the first low-latency, speaker-aware Sinhala synthesis pipeline to our knowledge.

II. 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 [2].

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 [3].

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 [8].

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

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 [10].

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 [11].

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 [12].

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 [13].

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 [14]

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 [15].

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 [16].

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 [17].

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 [18].

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, 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 [19]. 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 [20]. 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 [21].

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

III. Methodology

A. Fine-Tuning a Sinhala TTS Model

Although recent Sinhala text-to-speech (TTS) systems such as TacoSi [9] and the VITS-based Path Nirvana model [10] have demonstrated substantial improvements in speech quality and naturalness, they remain limited by their offline-only design and lack of speaker adaptation capabilities. These limitations restrict their use in real-time or personalized dubbing scenarios. To address this, we propose the fine-tuning of XTTS_v2, an open cross-lingual neural TTS architecture from Coqui [22] that supports real-time inference, zero-shot voice cloning, and emotion conditioning. XTTS_v2 is based on multilingual training, integrating speaker embeddings and attention-based synthesis modules, making it well-suited for under-resourced languages like Sinhala.

Our goal in fine-tuning XTTS_v2 is to enable natural and intelligible Sinhala speech synthesis while maintaining speaker identity and supporting real-time playback. To do this, we utilize the publicly available Path Nirvana dataset [5], which includes 11.8 hours of clean, phonetically diverse Sinhala speech recorded by a male speaker in a studio setting. The dataset underwent preprocessing involving silence trimming, punctuation-aware segmentation, normalization of numbers and symbols, and phoneme alignment. These steps helped stabilize the training process, reduced noise, and improved the model's prosodic variation. The quality and diversity of this dataset make it particularly suitable for fine-tuning a speaker-aware model capable of maintaining consistent voice identity across sentences.

B. Phase 1: Offline Dubbing Pipeline

In the first phase of the system, we implemente an offline pipeline designed for pre-recorded media content. This

pipeline follows a sequential architecture: automatic speech recognition (ASR), machine translation (MT), and Sinhala TTS synthesis. For ASR, we employ Faster-Whisper, an ONNX-optimized version of OpenAI's Whisper [7], which enables accurate transcription with significantly reduced latency and memory consumption [23]. It was selected for its excellent accuracy in English transcription and compatibility with real-time pipelines.

For machine translation, we integrate Meta's NLLB-200-1.3B model [24] using CTranslate2 for efficient Transformer inference. The NLLB-1.3B model offers a strong trade-off between translation fidelity and inference time, outperforming lightweight bilingual models in BLEU score and grammar preservation while still being deployable on modern GPUs. For Sinhala synthesis, we deploy a dual-path architecture: the primary method use our fine-tuned XTTS_v2 model, while a fallback option combined GPT-4o-mini-TTS with Seed-VC [25] for voice cloning. As this phase is not constrained by real-time requirements, we prioritize synthesis quality and identity consistency over latency.

C. Phase 2: Real-Time Dubbing Pipeline

The second phase focuses on enabling real-time English-to-Sinhala dubbing, such as during live video streaming or meetings. To achieve this, we propose to design a pipelined, sentence-level streaming architecture. Audio is first segmented using Silero VAD [26] and punctuation-aware boundaries to extract complete sentences. Each sentence chunk is then pass through three independent modules: ASR, MT, and TTS. While chunk n is being synthesized, chunk n+1 is translated, and chunk n+2 is being transcribed—thereby maintaining continuous flow with reduced cumulative delay.

We retain Faster-Whisper as our ASR module for its fast and robust transcription performance. Transcribed sentences are immediately feed into the MT module powered by NLLB-200-1.3B with CTranslate2. Despite its multilingual capacity, NLLB delivers fluent and morphologically correct Sinhala output with manageable latency when operated with sentence-level caching and batching.

The TTS stage follows a dual-path architecture, where users can choose between two synthesis approaches based on their needs. The primary option is our fine-tuned XTTS_v2 model, selected for its ability to produce natural, speaker-consistent Sinhala speech with low latency, particularly when deployed using ONNX or GPU acceleration. Alternatively, users may select a voice cloning path that combines GPT-4o-mini-TTS to generate initial speech, followed by Seed-VC for zero-shot voice conversion using a reference speaker audio [25]. This approach allows flexible integration of cloned voices without requiring per-speaker fine-tuning. Regardless of the selected TTS path, to synchronize the dubbed Sinhala audio with the original video's timing, we apply time-stretching techniques, which adjust playback speed without affecting pitch. This ensures smooth temporal alignment in live or semi-live scenarios.

IV. Conclusion

This work presents a modular and scalable real-time English-to-Sinhala dubbing system that addresses the key challenges of language accessibility, low-resource synthesis, and speaker identity preservation. By integrating open-source tools across each stage, we provide both an offline high-fidelity pipeline and a real-time streaming alternative. Our system is one of the first to support Sinhala dubbing with optional voice cloning and natural prosody, making it well-suited for personalized educational content, media localization, and low-cost accessibility solutions.

The project contributes meaningfully to under-resourced language technologies by demonstrating that real-time dubbing for Sinhala is both technically feasible and practically valuable using entirely open tools. Beyond addressing the technical limitations of current Sinhala TTS systems, our approach lays the groundwork for reusable, real-time speech translation pipelines for other morphologically rich, low-resource languages.

References

[1] Meta AI, "No Language Left Behind," arXiv:2207.04672, 2022.

[2] Weerasinghe et al., "Festival-si: A Sinhala Text-to-Speech System," SLTU, 2007.

[3] Nanayakkara et al., "A Human Quality Text to Speech System for Sinhala," MERCon, 2018.

[4] Kasthuriarachchi et al., "TacoSi: A Sinhala Text to Speech System with Neural Networks," ICIACS, 2023.

[5] Path Nirvana Foundation, "Sinhala TTS Dataset Release," GitHub, 2023.

[6] Meta AI, "NLLB-200: Multilingual Translation Model," arXiv:2207.04672, 2022.

[7] A. Radford et al., "Whisper: Robust Speech Recognition via Large-Scale Weak Supervision," OpenAI, 2022.

[8] De Silva N. "Survey on Publicly Available Sinhala Natural Language Processing Tools and Research", 2024

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

[10] Path Nirvana , "Sinhala VITS Neural TTS Model Release [GitHub/Technical Report]", 2023.

[11] Wijerathna, H. et al., "SEES: Sinhala to English and English to Sinhala Rule-based Translator. ICTer Conference", 2012.

[12] R. Perera, T. Fonseka, R. Naranpanawa, and U. Thayasivam , "Improving English to Sinhala Neural Machine Translation using Part-of-Speech Tag", 2022

[13] Wavel AI. (2023).

[14] VideoDubber. (2023).

[15] Kapwing TTS API. (2023).

[16] LingoCub Chrome Extension. (2024).

[17] Transmonkey. (2024).

[18] YouTube Dubbing – Translate & Dub Extension. (2024).

[19] OpenAI. (2024). GPT-4o Voice Capabilities.

[20] Google DeepMind. (2024). Gemini 2.5 Flash Technical Preview. https://deepmind.google/technologies/gemini

[21] Microsoft Azure. (2024). Text to Speech – Neural Voices: Sinhala.

[22] Coqui TTS Team, “XTTS-v2: Cross-lingual TTS with Voice Cloning,” GitHub, 2023.

[23] Faster-Whisper, GitHub Repository:

[24] Meta AI, “No Language Left Behind,” arXiv:2207.04672, 2022.

[25] Plachtaa, “Seed-VC: Real-Time Voice Conversion,” GitHub, 2023.