Survey on Real-Time English-to-Sinhala Speech Translation: A System-Level Review of Open Source Models and Pipelines

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

Abstract

Real-time speech translation from English to Sinhala involves transcribing English audio (ASR), translating the text (MT), and synthesizing Sinhala speech (TTS). Since Sinhala is a morphologically rich, low-resource language, building high-quality English–Sinhala translation is challenging. This survey reviews open-source solutions at each stage of the pipeline. We examine open ASR models for English recognition, open MT models for English-to-Sinhala translation, and open TTS systems for Sinhala speech synthesis. We compare their accuracy, latency, and resource requirements, and highlight recent research and existing prototype systems. Key gaps remain in data availability and end-to-end integration. We conclude with best practices for combining open components into an English-to-Sinhala speech translation pipeline, aiming to guide developers and researchers in this underserved language pair.

1 Introduction

Speech-to-speech translation enables real-time communication across languages. For English to Sinhala translation, the pipeline typically consists of English ASR to English text, followed by text MT, then Sinhala TTS. Sinhala (si) is the majority language of Sri Lanka, but is under-resourced in NLP. Its complex morphology and limited annotated data make robust translation difficult. Existing commercial tools (e.g. Google Translate) achieve only moderate quality (≈3.7/5 rating)[1] , and proprietary APIs may not be accessible to all users. This motivates an open-source approach: by combining freely available ASR, MT, and TTS models, one can build a translation system with no licensing cost. In this survey, we review open models for each stage, compare their performance (accuracy vs latency), and assess how they can be integrated in a real-time pipeline. We also discuss research on English–Sinhala MT and Sinhala TTS, and prototype systems that have been built.

2 Background and Task Definition

Real-time speech translation generally follows a cascade architecture: ASR (Speech-to-Text), MT (Text-to-Text), and TTS (Text-to-Speech). Each stage introduces errors, so overall quality depends on the weakest component. Metrics include word error rate (WER) for ASR, BLEU (or human scores) for MT, and Mean Opinion Score (MOS) or intelligibility for TTS. For English to Sinhala, additional challenges arise: English is SVO with relatively analytic morphology, whereas Sinhala is SOV and highly agglutinative [2]. Sinhala's verb conjugations and noun inflections make accurate translation hard without rich morphological knowledge. A direct end-to-end speech-to-speech model could be ideal, but no open end-to-end model exists for this pair; thus we assume a cascade pipeline.

3 Component Review

3.1 English ASR

Open ASR toolkits and models include OpenAI Whisper, Meta Wav2Vec2/XLSR, Kaldi/Vosk, Mozilla DeepSpeech, and SpeechBrain frameworks. OpenAI's Whisper is a pretrained Transformer-based model trained on 680k hours of multilingual audio [3]. Whisper supports English out-of-the-box (with English-only variants) and is widely considered the best open ASR today. Whisper comes in sizes (tiny→large) that trade off speed vs accuracy: the small model (244M params) is about 4× faster than large and uses 2 GB VRAM, while large (1.5B params) is 10× slower (1× relative speed). For real-time use on limited hardware, one would choose a smaller Whisper model or the "turbo" variant (which runs 8× faster than large with minimal quality loss [4].

Meta's wav2vec 2.0 / XLSR models are self-supervised speech encoders with strong ASR performance when fine-tuned. Pretrained English wav2vec2 models (e.g. facebook/wav2vec2-base-960h) achieve WERs on par with Whisper for clear speech, and XLS-R (trained on 436,000 hours of multilingual data) supports many languages. Kaldi is an open toolkit (C++) in which users train ASR pipelines [6]; recent Kaldi models for Sinhala achieved ≈28.6% WER [5], suggesting moderate accuracy. Vosk (based on Kaldi) provides ready-made English models (with WER 5–7% on LibriSpeech) that run offline and can stream with low latency [7]. DeepSpeech (Mozilla) and others are older and less accurate than Whisper or wav2vec-based models. In summary, Whisper offers very high English ASR accuracy (near-human for clear speech) with the trade-off of higher compute for large models [4]. Lower-latency options (e.g. Whisper-tiny/base, Vosk) run on lightweight devices but with higher WER.

3.2 English to Sinhala Machine Translation

When considering open source MT models supporting English to Sinhala, M2M-100 (Meta) is A many-to-many multilingual Transformer (12B params) that directly translates between 100 languages. It includes Sinhala (language code "si"), so one can use Helsinki-NLP/opus-mt-en-si or the meta/m2m model for English→Sinhala. M2M-100 was trained on massive parallel data and often yields better quality than pivoting.

Meta's NLLB-200 (No Language Left Behind) supports Sinhala and provides an open multilingual model (600M/1.3B/3.3B parameters) aimed at low-resource languages.

Helsinki-NLP/OPUS-MT project may have an English–Sinhala model (trained on OPUS web corpora), but coverage and quality are usually lower than M2M/NLLB.

Academic works have built English–Sinhala NMT using Transformer-based models (Fonseka et al., 2020, often adding subword or POS features. However, these models are not widely released for public use [8][9].

In human evaluations, commercial systems (Google, Bing) still outperform open MT: a recent study gave Google Translate an average rating of 3.7/5 and Bing 3.4/5 for English to Sinhala [1]. Among open/free methods, Meta's Llama LLM surprisingly scored 3.1/5 (close to Google) in that study [1], but dedicated MT models like M2M-100 or NLLB have not been independently evaluated in this context. In practice, M2M-100 and NLLB represent the best open options for translation quality, though their inference latency is significant (hundreds of milliseconds per sentence on CPU). Smaller distilled models or phrase-based systems will be faster but less accurate.

3.3 Sinhala TTS (Text-to-Speech)

Synthesizing high-quality Sinhala speech from text is challenging due to Sinhala's unique script and prosody. Early open efforts include Festival-si, a Festival-based diphone concatenation system. Festival-si achieved an intelligibility score of 71.5% in evaluation, but its voice sounds robotic [10]. Recent research has produced neural TTS for Sinhala. Notably, TacoSi (a Tacotron-based model) achieved a high Mean Opinion Score of 4.39 and 84% intelligibility, significantly outperforming older methods [11]. Open frameworks like Coqui TTS or Mozilla TTS can train Sinhala voices if data is available.

A major asset for Sinhala TTS is the Path Nirvana TTS Dataset: 13.7 hours of multi-speaker speech (mixed male/female) with paired Sinhala text [12]. This dataset (GPL-3.0 licensed) has enabled modern neural TTS (e.g. VITS models) for Sinhala. In contrast, earlier work had only 7.5 hours of speech and no public fine-tuned models [12]. For voice cloning, open-source toolkits (e.g. "Real-Time Voice Cloning") could adapt to Sinhala speakers given reference audio, but to our knowledge no specialized Sinhala cloning model exists. Time-stretching (speed adjustment) is an optional post-processing to match synthesis speed or fit time constraints, but adds complexity.

4 Comparative Review of Tools

Comparing ASR Tools, Whisper is large, accurate, slower model while its small versions are faster. Wav2Vec2/XLSR requires fine-tuning, but fast inference once done. Vosk/Kaldi is accurate ¡10% WER on English, can run on edge devices with small models [6][7]. DeepSpeech is currently outdated, with limited record length. As a conclusion Whisper-large gives best accuracy (English WER 2% on clean speech) but needs GPU; Whisper-tiny or Vosk small can do real-time transcription on CPU with higher WER. But as a solution for real-time / near real-time you can, Faster-Whisper: an optimized implementation of OpenAI's Whisper ASR model, built using ONNX and CTranslate2 backends. It delivers real-time inference speeds on both CPU and GPU with minimal loss in transcription quality compared to the original Whisper models [1]. Combined with Silero VAD for voice activity detection, it becomes ideal for streaming ASR pipelines. Its support for chunked transcription and low latency makes it a suitable frontend for real-time dubbing systems [13].

MT Tools such as M2M-100/NLLB are state-of-the-art but heavy (hundreds of MB models). Their quality on Sinhala is likely higher than any purely open bilingual model. Helsinki/OPUS MT models (if available) are lightweight (300M params) but have lower BLEU [14] . In terms of latency, transformer MT on CPU is on the order of 100–300 ms per sentence; caching or smaller distilled models can speed this. While larger variants like NLLB-3.3B offer superior translation accuracy, the 1.3B version provides a strong balance between translation quality and inference latency [15], especially when used with CTranslate2 for fast deployment. It supports direct English to Sinhala translation and has been validated in studies as outperforming previous open models in quality.

In Text-To-Speech perspective Festival-si (runs in real time on CPU, but naturalness is low [10]. Neural TTS (Tacotron, FastSpeech, VITS) produce far more natural output but need GPU for training and often for fast inference. For example, a Tacotron model like TacoSi yields MOS 4.39 [11], while Festival-si's intelligibility was only 71.5% [10]. With the Path Nirvana dataset (13.7h), one can train a high-quality Sinhala neural TTS (coqui TTS or SpeechT5) that rivals proprietary voices. Voice cloning toolkits require a pretrained synthesizer; one could fine-tune a Sinhala TTS on a specific speaker's data. Time-stretching libraries can adjust output speed but introduce latency. However, neither supports real-time streaming or voice cloning. Fine-tuning XTTS_v2, a multilingual cross-lingual TTS model, on Sinhala data (like Path Nirvana) is a promising path to enable real-time, speaker-aware Sinhala TTS.

5 Existing Systems

Few end-to-end English–Sinhala speech translation systems have been documented. In 2023, students from SLIIT developed a prototype English–Sinhala speech translation system using open-source ASR, MT, and TTS components. Their system showcased the feasibility of open-source Sinhala speech pipelines for constrained domains [16]. However, it lacked real-time streaming and speaker identity preservation. There are also several other systems relied on Google APIs and are not open.

6 Gaps and Analysis

Data scarcity is a major gap. Sinhala–English parallel corpora exist but quality issues were reported and they are small relative to high-resource pairs. Monolingual Sinhala text and speech corpora have grown recently, and the Path Nirvana speech data [17] but remain limited for large models. The lack of large

clean training sets hinders both ASR and NMT for Sinhala.

Open models like Whisper and M2M are computationally heavy. Running Whisper-large or NLLB on CPU can exceed real-time constraints for long utterances. Smaller models exist but at a quality cost. Real-time end-to-end throughput also depends on TTS speed; neural TTS can have latency of several hundred ms per utterance. Efficient inference (quantization, on-device acceleration) is needed for practical real-time use.

Errors compound across stages. ASR errors degrade MT input, and MT errors (due to out-of-vocabulary or grammar) yield unnatural Sinhala text. Colloquial or code-switched English complicates ASR/MT. Sinhala's morphological richness means simple direct translation often misses nuances. For example, morphological generation is needed in TTS to convert Roman script or transliterations into proper Sinhala orthography. Existing MT models may transliterate names and numbers poorly. Pipeline integration (audio chunking, punctuation insertion, voice cloning) is non-trivial.

Despite progress, no Sinhala TTS model currently supports both real-time synthesis and speaker identity preservation. While TacoSi and Path Nirvana produce high-quality speech, they are slow or single-speaker only [11] [17]. There is no open Sinhala TTS that can stream output on-the-fly or adapt to different speakers. Voice cloning toolkits offer the necessary mechanisms, but require Sinhala-specific fine-tuning. Developing a fine-tuned XTTS_v2 Sinhala model with real-time inference support remains one of the most impactful next steps.

In summary, while open ASR, MT, and TTS models exist individually, a turnkey open-source English to Sinhala speech translation system is still a research prototype. Key gaps are training data and system engineering.

7 Conclusion

In this review, we surveyed open-source components for building an English to Sinhala speech translation pipeline. For ASR, OpenAI's Whisper (multilingual model) offers state-of-the-art accuracy, with trade-offs in latency between model sizes [4]. For MT, Meta's M2M-100 or NLLB models (open) currently provide the strongest English–Sinhala translation performance [14], though commercial systems still rate slightly higher. For TTS, concatenative systems like Festival (free) can run in real time but sound robotic [10], whereas neural TTS (e.g. Tacotron) yields human-like Sinhala with MOS≈4.4 [11]. Researchers have created Sinhala TTS datasets (e.g. 13.7 h multi-speaker) to support these model [12]. Existing demonstrations show that English to Sinhala voice translation is feasible, but a polished open pipeline remains an engineering challenge. The main gaps are limited high-quality Sinhala corpora and the need for real-time optimization. Future work should focus on curating more data, fine-tuning open models for Sinhala, and integrating the stages into a low-latency system. With these advances, a fully open, accurate English-to-Sinhala speech translator could become widely accessible to serve millions of Sinhala speakers.

References

[1] R. Jayakody and G. Dias, *Performance of Recent Large Language Models for a Low-Resourced Language*, 2024.

[2] A. Pramodya, *Exploring Low-resource Neural Machine Translation for Sinhala-Tamil Language Pair*.

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

[4] Whisper GitHub, [Online]. Available: `https://github.com/openai/whisper`

[5] B. Gamage, R. Pushpananda, T. Nadungodage, and R. Weerasinghe, *Improving Sinhala Speech Recognition Through e2e LF-MMI Model*, 2021.

[6] Kaldi Speech Recognition Toolkit, [Online]. Available: https://kaldi-asr.org

[7] Vosk, *Offline Speech Recognition Toolkit*, [Online]. Available: https://alphacephei.com/vosk/

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

[9] Aaivu-machine-trans-eng-sin, [Online]. Available: https://github.com/aaivu/aaivu-machine-trans-eng-sin

[10] R. Weerasinghe, A. Wasala, V. Welgama, and K. Gamage, *Festival-si: A Sinhala Text-to-Speech System*, 2007.

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

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

[13] CTranslate2, GitHub, 2023. [Online]. Available: https://github.com/OpenNMT/CTranslate2

[14] Huggingface, [Online]. Available: https://huggingface.co/

[15] Meta AI, *No Language Left Behind: Scaling Human-Centered Machine Translation*, arXiv:2207.04672, 2022.

[16] SLIIT NLP Group, *Sinhala-English Voice Translator Project*, Student Project Report, 2023.

[17] Path Nirvana Foundation, *Sinhala TTS Dataset Release*, GitHub, 2023. [Online]. Available: https://github.com/pathnirvana/sinhala-tts-dataset