Training a Sinhala Speech-to-Text Model: Data, Training and Results

Paper | 2026-09-08 | Seniru Epasinghe, Tharindu Sathsara, Dilanka Kasun, Sahan Dileepa, Bhawanthi Weerasinghe

How we trained Chakra's Sinhala speech-to-text on public recordings plus data we built and verified ourselves, cut the error rate, and tested it on real spontaneous speech.

Abstract

Sinhala is a low-resource language, and general speech models handle it poorly, especially on phone audio. We trained Sinhala speech-to-text models on a public set of 161,296 Sinhala clips together with datasets that we collected and verified ourselves, end to end, with our own team. In early experiments on the public set, the better of two training runs reached 23.6% word error rate (20.4% with Sinhala-safe scoring). Our first model, prepared by Seniru Epasinghe and trained on about 100,000 clips from many speakers in nearly 30 hours on L40 GPUs, was followed by further rounds of training. Our current production models were then tested on 49 clips of real spontaneous speech from outside their training data.

Key findings

- Training data = a public set of 161,296 Sinhala clips (16 kHz mono; 145,166 for training and 16,130 for testing) plus datasets we built ourselves: audio collected from public sources, then checked end to end by our team
- Early experiments on the public set: run 1 (2 epochs) 26.72% word error rate; run 2 (3 epochs, higher learning rate, cosine schedule) 23.60% raw and 20.40% with Sinhala-safe scoring
- Learning flattened out at about 1.8 epochs: more epochs will not help, more and cleaner data will
- The usual open-source text normaliser deletes Sinhala vowel signs and so gives misleading scores; we built a Sinhala-safe one
- The first model used about 100,000 clips (10 to 30 seconds each, from many speakers) and took nearly 30 hours to train on L40 GPUs; further rounds of training followed
- On 49 clips of real spontaneous speech they never trained on: 11.8% character error rate (larger model) and 14.4% (smaller model)
- Sinhala needs about 17.8 model tokens per second of speech, so the usual 128-token output limit cut callers off after about 7 seconds; scaling the limit to the audio length took one model from 76.7% to 39.9% word error rate without retraining

1. Why Sinhala needs its own model

Most speech models are trained mainly on English and a handful of large languages. Sinhala has far less public data, a script where one vowel sign changes the whole word, and phone audio that is narrow-band and noisy. For a phone helpline such as Govimithuru, a wrong transcript means a

wrong answer.

2. Data

We combined two kinds of data.

- **A public Sinhala speech dataset** of 161,296 clips, 16 kHz mono, split 90/10 with a fixed seed into 145,166 training and 16,130 test clips. Audio is decoded on the fly, because pre-computing the audio features for all clips needed about 145 GB of disk.
- **Our own datasets.** We collected audio from public sources and verified it end to end with our own team: every clip checked against its transcript. Team member Seniru Epasinghe prepared further datasets for the project.

3. Training

We started from a strong open speech model and trained it for Sinhala. Our first model, prepared by Seniru Epasinghe, used about 100,000 clips of 10 to 30 seconds from many speakers and took nearly 30 hours to train on L40 GPUs. Later rounds of training added more data. The comparison below comes from experiments on a single 48 GB GPU.

- **Run 1:** 2 epochs, learning rate 5e-6, linear schedule.
- **Run 2:** 3 epochs, learning rate 1e-5, cosine schedule, effective batch size 16, mixed precision, gradient checkpointing.

4. Scoring Sinhala fairly

Standard text normalisers remove all Unicode 'mark' characters, which in Sinhala are the vowel signs. That makes wrong words look right. We score with a Sinhala-safe normaliser that removes punctuation and symbols but keeps the vowel signs.

5. Results

Run	Epochs	Word error rate (raw)	Word error rate (Sinhala-safe)
1	2	26.72%	-
2	3	23.60%	20.40%

The best checkpoint came at about 2.65 epochs. Error rate fell quickly during the first two epochs and barely changed after about 1.8.

6. From the lab to phone calls

Our current production models were tested on 49 clips of spontaneous Sinhala speech they had never seen. Character error rate is the fairer measure for Sinhala, because word error rate counts a whole word wrong for a single vowel sign.

Model	Character error rate	Word error rate
Smaller model	14.4%	39.9%

Larger model	11.8%	32.7%

A hidden problem: Sinhala costs about 17.8 tokens per second of speech, so a fixed 128-token output limit silently cut every caller off after about 7 seconds and returned a clean-looking first sentence. Scaling the limit to the length of each recording fixed this and removed looping outputs.

7. Limits and next steps

The 49-clip test is small, and the training data is not telephone audio. The next steps are more and cleaner data, telephone-style augmentation, and a larger model.