Teaching a Machine to Speak Sinhala: Training a Text-to-Speech Voice

Paper | 2026-09-26 | Seniru Epasinghe, Tharindu Sathsara, Dilanka Kasun, Sahan Dileepa, Bhawanthi Weerasinghe

How we built a clear Sinhala voice for phone calls, and the speed-versus-clarity test that set it to 16 generation steps.

Abstract

A phone assistant is only as good as its voice. We trained a neural text-to-speech model for Sinhala and chose its settings by measurement. We tested clarity by synthesising 60 real replies from live calls and having our own speech-to-text read them back. At 16 generation steps the voice scored a 4.9% character error rate, lower than real callers (11.8%), while generating speech about 1.5 to 2 times faster than at 32 steps.

Key findings

- Clarity test: 60 real agent replies, synthesised and read back by our Sinhala speech-to-text
- 16 steps: 4.9% character error rate, 5.3-7.0 seconds of speech generated per second. 32 steps: 5.0%, 3.5 seconds per second. 12 steps: 5.6%. 8 steps: 12.1%
- Through the full production path the voice scored 6.5% character error rate (real callers: 11.8%)
- Best piece length is 40-150 characters; shorter pieces can lose their first word
- Sinhala letters only: English words are silently dropped, so numbers and units are spoken as words
- Splitting speech-to-text and text-to-speech onto separate cards nearly tripled the number of simultaneous calls

1. The goal

A warm, clear Sinhala voice that is fast enough for live phone calls, and that real farmers can understand the first time.

2. Approach

We trained a neural text-to-speech model for Sinhala on the same data we collected for speech-to-text: recordings verified end to end by our own team, together with datasets prepared by Seniru Epasinghe. Training took more than 30 hours on L40 GPUs. A new voice is added with a short reference recording and its exact transcript, so changing the voice does not need retraining the whole model.

3. Measuring clarity without asking listeners

We took 60 real agent replies from live calls, synthesised them, and had our own Sinhala speech-to-

text transcribe the audio. A clear voice should be transcribed as accurately as a clear human speaker. For comparison, real callers score 11.8% character error rate.

Generation steps	Character error rate	Word error rate	Speech generated per second
32	5.0%	15.1%	3.5 s
16 (chosen)	**4.9%**	**15.3%**	5.3-7.0 s
12	5.6%	16.9%	7-9 s
8	12.1%	-	8.6 s

Sixteen steps is as clear as 32 and much faster. At eight steps, errors double.

4. Rules we learned

- Send 40-150 characters at a time, split at sentences or commas.
- Write the text the way it should be spoken: numbers, units and prices as words, with no emoji, lists or links.
- The voice speaks Sinhala letters only; English words are dropped.

5. Running it at scale

On one graphics card, speech-to-text and text-to-speech compete for the same chip. Putting each on its own card raised the comfortable number of simultaneous calls from about 3 to about 8 (see the voice pipeline article). The voice itself did not change: 6.5% character error rate through the full path, against 6.4% on a single card.

6. What is next

More voices, more natural pauses and expression, and a Tamil voice.