Recording Farmers' Voices: A Data Collection Standard for Sinhala Agricultural Speech

White paper | 2026-09-29 | Seniru Epasinghe, Tharindu Sathsara, Dilanka Kasun, Sahan Dileepa, Bhawanthi Weerasinghe

The recording and transcription rules we set to collect clean, natural farmer speech, from silence padding to how numbers are written.

Abstract

Good speech models start with good data. To teach a system to understand how farmers really talk, with regional accents, farming vocabulary and natural hesitations, we defined a strict collection and transcription standard. This paper explains each rule and why it is there.

Key findings

- Clips of 10-30 seconds, with exactly 1 second of silence before and after the speech
- Lossless WAV or FLAC, later standardised to 16 kHz mono; exactly one speaker per recording
- 50% male and 50% female voices, balanced across three age groups (18-35, 36-55, 55+) and across regional accents (for example Southern, Rajarata, Upcountry)
- Natural farm background noise allowed; conversations, traffic, music and echoey rooms are not
- Speech is elicited with scenarios (for example 'ask about urea prices'), not scripts
- Transcripts are verbatim, with stutters and filler words kept, no grammar fixes, and numbers written as Sinhala words

1. Why a standard

A phone line for farmers must understand people who speak in different accents, in fields, with farming words and prices. Public datasets are mostly read speech from quiet rooms, so we wrote our own rules for new recordings.

2. Audio

- Each clip is 10 to 30 seconds long, never more than 30.
- Exactly 1 second of silence before the speech starts and after it ends.
- Lossless WAV or FLAC; the pipeline converts to 16 kHz mono.
- One speaker per recording. Overlapping speech is not allowed.

3. Environment

Light wind, distant tractors and bird sound are allowed, since that is where farmers call from. Background conversation, loud traffic, music and indoor echo are not.

4. Speakers

The final dataset must have 50% male and 50% female speakers, a balanced split of young adults, middle-aged and senior adults, and natural regional accents. We do not standardise pronunciation.

5. Content

Recordings cover crop identification and diseases, agrochemicals, fertilisers and brand names, and prices and land measures such as akkaraya and parchasa. Speakers answer scenarios instead of reading scripts, for example 'ask about the price of urea'.

6. Transcription

One CSV maps each audio file name to its text. Transcripts are verbatim: stutters and fillers stay, the speaker's grammar mistakes are not corrected, and numbers are written as Sinhala words, not digits.

7. Conclusion

These rules keep the data natural, balanced and consistent, which is what a speech model needs to work for every farmer, not just the ones who sound like a news reader.