Sukuma Text-to-Speech

Zephania Reuben Isack Odero

Technical report · August 2026 · Research preview

Quality SKU: https://huggingface.co/nileagi/nileagi-suk-tts

Lite SKU: https://huggingface.co/nileagi/nileagi-suk-tts-lite

Abstract

This report describes NileAGI's two public Sukuma text-to-speech systems, nileagi/nileagi-suk-tts (quality) and `nileagi/nileagi-suk-tts-lite` (lite), both of which synthesize a single male read-aloud voice at 24 kHz from Sukuma text written in a Latin orthography that preserves vowel macrons. Both models were fine-tuned from a shared neural codec recipe on approximately 27.8k CTC-aligned audio clips, and the training audio itself is not redistributed together with the checkpoints. Quality is assessed through human listening on short Sukuma lines rather than an automatic score. Input to both systems must already be Sukuma text: anyone starting from Swahili should translate with `nileagi/nileagi-suk-mt` first, since neither TTS model accepts Swahili or English text directly.

1 Introduction

Sukuma (Kisukuma, ISO 639-3 `suk`) is a Bantu language spoken by several million people across the Lake Victoria region of northern Tanzania. As with other Bantu languages of the region, its grammar includes extensive noun-class marking and verb agreement, features that most existing TTS research and tooling were never built around, since that work has focused overwhelmingly on higher-resource, non-Bantu languages. Neural speech synthesis for Sukuma is essentially nonexistent in the public domain, and the systems described in this report are meant as an early step toward closing that gap rather than a finished product. NileAGI publishes two operating points that share a single, narrowly defined task: Sukuma text goes in, Sukuma speech comes out, at a fixed 24 kHz output rate and in the same macron-based orthography used across NileAGI's other Sukuma tools.

Table 1: Public TTS SKUs.

Role	Hub id	Selection rule
Quality	`nileagi/nileagi-suk-tts`	best listening quality
Lite	`nileagi/nileagi-suk-tts-lite`	same recipe, smaller footprint

Both models belong to the collection https://huggingface.co/collections/nileagi/nileagi-suk, alongside NileAGI's Sukuma speech-to-text and machine translation releases, and together the three systems are intended to compose into a full spoken Swahili-to-Sukuma pipeline.

2 Task

Direction. The supported mapping is Sukuma text to Sukuma speech at 24 kHz. Neither English nor Swahili is accepted as direct TTS input, and feeding either language's text to these models will not produce reliable speech, since the grapheme-to-sound mapping the models learned is specific to Sukuma.

Orthography. Input text uses Latin script with macrons (`ā ē ī ō ū` preserved throughout). Dropping the macrons from input text changes the pronunciation the model produces, since the macron carries real phonological information rather than being a purely cosmetic mark, so users who strip diacritics before synthesis should expect noticeably different output than what is reported here.

Voice. This release starts with a single male read-aloud voice, identified internally as `nileagi_suk`, baked into each checkpoint rather than being selectable at inference time. This is a deliberate starting point rather than a ceiling: multi-speaker support, including voice cloning and a female voice option, is the natural next step, and we intend to extend to multiple speakers once suitable reference recordings become available. Neither is supported in this first release.

3 Training material

Supervision for both SKUs comes from CTC force-aligned Sukuma read-aloud clips, using the same style of alignment pipeline described in NileAGI's Sukuma speech-to-text report. Only alignments explicitly tagged `ctc` are kept for training, and each training row pairs one audio clip with its Sukuma transcript together with a fixed reference prompt clip used for speaker embedding during fine-tuning.

Table 2 summarises the fine-tuning inventory shared by both SKUs, including model backbone sizes and the small handful of settings that differ between them.

Table 2: Fine-tuning inventory (both SKUs).

Quantity	Value
Training clips	27,827
Sample rate (output)	24 kHz
Reference prompt	one clean, mid-length clip
Quality backbone size	∼1.7B parameters
Lite backbone size	∼0.6B parameters
Fine-tuning epochs	3 (each SKU)
Speaker id	`nileagi_suk` (codec slot 3000)

Training audio is deliberately not published alongside the model weights, consistent with NileAGI's other Sukuma releases, though one short demonstration WAV file (`samples/sample_04.wav`) ships with each Hub repository so that anyone can listen to representative output before committing to a full inference setup.

4 Systems

Both SKUs are neural TTS models fine-tuned from a shared codec-based recipe that operates at a 12 Hz frame rate. In this setup, speech is represented as a sequence of discrete audio tokens produced by a neural codec, rather than as a continuous waveform or spectrogram predicted directly. Inference always uses the fixed voice id `nileagi_suk` together with a language setting of `Auto`, and short sentences are strongly preferred over long paragraphs, which should be split before synthesis for the most reliable output.

- **Quality** (`nileagi-suk-tts`): the larger of the two backbones, offering noticeably higher naturalness on short to medium-length lines.
- **Lite** (`nileagi-suk-tts-lite`): a smaller backbone that trades some naturalness for lower VRAM requirements and faster synthesis.

5 Evaluation

The primary evaluation method for this release is human listening rather than an automatic metric, following standard subjective evaluation practice for speech synthesis quality. On short Sukuma lines, listeners can generally expect the following pattern of behaviour:

- a clear, consistent male read-aloud timbre on short-to-medium sentences;
- audible pronunciation drift whenever input macrons are omitted;
- weaker and less predictable behaviour on long or conversational prompts that depart from the literary read-aloud register the models were trained on;
- the lite SKU sitting somewhat below the quality SKU in perceived naturalness, consistent with its smaller backbone.

No automatic mean opinion score, WER, or CER figure is reported for either SKU in this release, and any such automatic metric would in any case need careful validation against human judgement before it could be trusted for this language and voice.

6 Deployment

Both SKUs are distributed as standard Hugging Face model repositories, and either can be loaded directly from the Hub for local inference or downloaded for offline use. The quality checkpoint requires approximately 8 GB of free GPU memory at inference time, while the lite checkpoint is intended for environments where that footprint is not available.

Listing 1 shows the minimal steps for loading the quality SKU from the Hub and synthesising a short Sukuma sentence; the lite SKU can be used in exactly the same way by substituting its Hub id.

Listing 1: Loading the model from Hugging Face (Python)

```python
from transformers import AutoModel, AutoProcessor

```

```
repo = "nileagi/nileagi-suk-tts"
processor = AutoProcessor.from_pretrained(repo)
model = AutoModel.from_pretrained(repo)

inputs = processor(text="Ulili mhola?", voice="nileagi_suk")
audio = model.generate(**inputs)
processor.save_audio(audio, "out.wav")
```

7 Intended use and non-goals

Intended: research-oriented speech synthesis; short narrative lines in the literary read-aloud register the models were trained on; the final hop of a pipeline that goes from Swahili text to Sukuma speech; offline demonstrations of Sukuma speech technology.

Not intended (for now): direct English or Swahili text-to-speech input; conversational or singing speech generation; commercial products deployed without a licence from NileAGI. Multi-speaker voice cloning and a female voice are not part of this first release, but they are on the roadmap rather than ruled out.

8 Licence

The weights for both SKUs are released under CC BY-NC-SA 4.0. Organisations that need commercial licensing terms should reach out directly at mailto:hi@nileagi.com.

9 Conclusion

NileAGI releases two Sukuma text-to-speech SKUs, fine-tuned from a shared codec-based recipe on roughly 27.8k CTC-aligned read-aloud clips of a single male voice. The quality SKU is the right default whenever naturalness matters most and the required GPU memory is available; the lite SKU is the better fit when footprint or latency is the binding constraint. As with NileAGI's other Sukuma releases, both systems represent an early research preview built on a narrow literary register and a single voice, rather than a general-purpose Sukuma speech synthesis product. Quality in this report is judged by ear rather than by an automatic score, and that human-listening protocol should be kept in mind when interpreting any claims made here.

10 Citation

Listing 2: BibTeX entry

```
@misc{nileagi-suk-tts-2026,
  title        = {Sukuma Text-to-Speech: Technical Report},
  author       = {Zephania Reuben and Isack Odero},
  year         = {2026},
  howpublished = {Hugging Face},
  url          = {https://huggingface.co/nileagi/nileagi-suk-tts}
```

```
}
```

References

[1] A. Graves, S. Fernández, F. Gomez, J. Schmidhuber. *Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks.* ICML, 2006. — the CTC objective used to force-align read-aloud clips to their Sukuma transcripts in Section 3.

[2] V. Pratap et al. *Scaling Speech Technology to 1,000+ Languages.* arXiv: 2305.13516, 2023 (JMLR, 2024). — background on multilingual CTC-based forced alignment at scale, the general technique behind the alignment step in Section 3.

[3] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, M. Sonderegger. *Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi.* Interspeech, 2017. — a widely used reference implementation of CTC-based forced alignment for read-aloud speech.

[4] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, M. Tagliasacchi. *SoundStream: An End-to-End Neural Audio Codec.* IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021. — background on the discrete neural audio codec representation underlying the shared 12 Hz codec recipe in Section 4.

[5] A. Défossez, J. Copet, G. Synnaeve, Y. Adi. *High Fidelity Neural Audio Compression.* arXiv:2210.13438, 2022. — further background on neural audio codecs of the kind used to represent speech tokens in Section 4.

[6] C. Wang et al. *Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers.* arXiv:2301.02111, 2023. — background on codec-token language modelling for TTS and on using a short reference clip for speaker conditioning, the general approach reflected in the reference-prompt mechanism described in Section 3.

[7] A. van den Oord et al. *WaveNet: A Generative Model for Raw Audio.* arXiv:1609.03499, 2016. — foundational background on neural generative models for raw speech audio.

[8] J. Kong, J. Kim, J. Bae. *HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis.* NeurIPS, 2020. — background on GAN-based neural vocoders used for high-fidelity waveform generation in modern TTS pipelines.

[9] ITU-T. *Recommendation P.800: Methods for Subjective Determination of Transmission Quality.* International Telecommunication Union, 1996. — standard methodology underlying human-listening evaluation of speech quality, the primary evaluation approach used in Section 5.

[10] D. P. Kingma, J. Ba. *Adam: A Method for Stochastic Optimization.* ICLR, 2015. — the optimizer family underlying the fine-tuning procedure used for both SKUs.