WAXAL: Google Releases 2,400 Hours of Speech Data for 27 African Languages

Kabui, Charles

2026-03-19

Read at ToKnow.ai

Google Releases 2,400 Hours of Speech Data for
27 African Languages
WAXAL: open ASR and TTS dataset for 100M+ speakers across Sub-Saharan Africa
27
Languages across 26+ countries
1,846 hrs
Natural speech for ASR (unscripted)
565 hrs
Studio-quality TTS recordings
March 19, 2026
ToKnow.ai

Google Research has released WAXAL, a large-scale open speech dataset covering 27 Sub-Saharan African languages spoken by over 100 million people across 26+ countries. Funded by Google and the Gates Foundation, the dataset has two parts: WAXAL-ASR, with roughly 1,846 hours of transcribed natural speech, and WAXAL-TTS, with over 565 hours of studio recordings for voice synthesis. The ASR data uses an image-prompted method where speakers describe visual scenes in their native language rather than reading scripts, capturing authentic tonal patterns and code-switching (naturally switching between languages mid-sentence). The TTS data was recorded by community members who drafted scripts and recorded in pairs, with some building custom studio boxes using project funding.

What makes WAXAL unusual is who built it. African institutions led all data collection: Makerere University handled nine languages, the University of Ghana covered eight, and Digital Umuganda, Media Trust, Loud n Clear, and AIMS Senegal contributed the rest. Partners retain ownership of their data. Everything is released under CC-BY-4.0 on HuggingFace. For anyone building voice assistants or transcription tools in languages like Wolof, Luganda, or Amharic, this is the largest openly licensed resource available. Cohere's Tiny Aya showed on-device multilingual models are possible; WAXAL provides the speech data to extend that reach to voice.

Over 2,000 languages are spoken in Sub-Saharan Africa, and almost none have adequate speech technology support. WAXAL does not solve that, but it proves a replicable model: fund local institutions, let them lead collection with shared methodology, and release the data openly. The paper is already spawning derivative research, including fine-tuned Whisper models and benchmarks for 13 African languages.

Sources:

- Google Research Blog: WAXAL
- WAXAL Paper (arXiv)
- WAXAL Dataset on HuggingFace
- Makerere University
- Digital Umuganda

Disclaimer: *For information only. Accuracy or completeness not guaranteed. Illegal use prohibited. Not professional advice or solicitation.* ***Read more:*** ***/terms-of-service***