

Kayode Olaleye

Data quality and evaluation for low-resource and multilingual speech and language processing
kaykola.olaleye@gmail.com · Ibadan, Nigeria · [Google Scholar](#) · [ACL Anthology](#) · [LinkedIn](#)

PROFILE

Speech and language researcher working on how multilingual and low-resource language data is built, validated and evaluated. Co-Investigator on a Gates Foundation programme that released a 3,000-hour speech corpus across seven South African languages, and first author of a code-switched dataset across five languages corrected by native speakers. Google PhD Fellow (2019).

SELECTED PUBLICATIONS

AfroCS-xs: Creating a Compact, High-Quality, Human-Validated Code-Switched Dataset for African Languages. Olaleye et al. ACL 2025, Long Papers. *First author. LLM-generated code-switched text across Afrikaans, Sesotho, Yoruba, isiZulu and English, corrected by native speakers; automatic metrics diverged from speaker judgement.*

Swivuriso: The South African Next Voices Multilingual Speech Dataset. Marivate, Olaleye et al. 2025, arXiv:2512.02201. *Second author. Corpus design, annotation guidelines, quality control across seven language teams.*

Keyword Localisation in Untranscribed Speech Using Visually Grounded Speech Models. Olaleye, Oneata, Kamper. IEEE JSTSP 2022, vol. 16. *First author.*

YFACC: A Yoruba Speech-Image Dataset for Cross-Lingual Keyword Localisation through Visual Grounding. Olaleye, Oneata, Kamper. IEEE SLT 2022. *First author. Collected and released the corpus.*

Attention-Based Keyword Localisation in Speech Using Visual Grounding. Olaleye, Kamper. Interspeech 2021. *First author.*

DATASETS RELEASED

Swivuriso (2025) — 3,000+ hours, 2,353 speakers, seven South African languages, CC BY 4.0. 30+ ASR models fine-tuned on it; underpins Microsoft's Paza-bench.

AfroCS-xs (2025) — human-validated code-switched data, five languages.

YFACC (2022) — Yoruba spoken-caption corpus extending Flickr8k. Sole collector.

EXPERIENCE

Postdoctoral Research Fellow, Data Science for Social Impact, University of Pretoria 2023–2025

- Co-Investigator, African Next Voices South Africa (Gates Foundation); contributor to the funding proposal.
- Work package lead for corpus quality assurance: seven language teams, 1,000+ contributors, owner of the transcription and annotation quality check.
- Built consent-based collection protocols restricting downstream voice cloning. Co-supervised one PhD and one MSc candidate.

Independent Researcher, Speech and Language Processing 2025–present

- ASR, LLM evaluation and speech data for low-resource African languages. Quantised edge ASR for local dialects; offline-first pipelines for low-bandwidth settings.

EDUCATION

PhD, Electronic Engineering (speech processing), Stellenbosch University, 2023. Supervisor: Prof Herman Kamper. Thesis: Towards Visually Grounded Speech Models for Keyword Localisation.

MSc, Mathematical Sciences, Stellenbosch University, 2018. Structured Master, African Institute for Mathematical Sciences, 2016.

BTech, Industrial Mathematics, Federal University of Technology Akure, 2012.

SELECTED AWARDS AND SERVICE

Google PhD Fellowship (2019). DAAD In-Region Scholarship (2019–2022). AIMS-NEI Scholarship.

Co-organisier, AfricaNLP Workshop 2024. Reviewer: ACL Rolling Review, ML-SUPERB Challenge (Interspeech 2023), ReMuCAI (NeurIPS 2026 Workshop).