Voices of Fashion: 150-Hour Yoruba Speech Corpus for Egba Adire and Color Advisory Systems

Abstract

The intersection of language, culture, and fashion offers a unique opportunity to design inclusive artificial intelligence (AI) systems that address underrepresented communities. This work focuses on the curation of a **150-hour Yoruba fashion speech dataset** for the Egba people of Southwestern Nigeria, a subgroup of the Yoruba estimated at nearly 2 million speakers in Ogun State. The Egba are historically renowned for their mastery of *Adire* (tie-and-dye) textile craft, with Abeokuta recognized as the capital of Nigeria's Adire industry and home to Itoku Market, the largest tie-and-dye marketplace in West Africa. Despite this cultural and economic significance, Egba communities remain largely excluded from AI-driven tools that could enhance creativity, advisory, and digital commerce in their local language.

We propose the collection of speech recordings involving **200+ Egba speakers** across artisans, market women, designers, and young customers. The dataset will capture fashion-related expressions, fabric pattern terminologies, dyeing techniques, color combinations, and advisory dialogues in Yoruba with Egba dialectal variations. All recordings will be transcribed, annotated, and validated by Yoruba language experts to ensure both **linguistic richness and cultural accuracy**.

The resulting dataset will enable localized AI applications, such as a Yoruba-speaking fashion advisory chatbot, capable of recommending designs, color blends, and tie-and-dye techniques in culturally grounded ways. Beyond advisory, the dataset supports **automatic speech recognition (ASR)**, **machine translation**, and **dialogue systems** tailored to Yoruba-speaking populations. By situating this work at the intersection of low-resource language dataset curation, AI for cultural heritage, and digital empowerment, we aim to directly contribute to **economic inclusion, cultural preservation, and sustainable innovation** for one of West Africa's most historically influential communities.

Table 1. Overview of Dataset Specifications

Feature	Specification
Total Duration	150 hours
Speakers	200+
Dialect Coverage	Yoruba (Egba variant)
Domains	Fashion advisory, color naming, textile patterns, market dialogues
Annotations	Transcription + Translation + Cultural tags
Applications	ASR, MT, Chatbots, Fashion AI

References

[1] Alexandra *et al*. "Developing an Open-Source Corpus of Yoruba Speech," 2020.
[2] Bird, S. "Natural Language Processing for Low-Resource Languages." *Computational Linguistics*, 2020.
[3] Bamgbose, A. "Language and the Nation: The Language Question in Sub-Saharan Africa." Edinburgh University Press, 1991.

Keywords

Yoruba speech dataset, low-resource language technology, ASR, multilingual NLP, dataset curation, cultural heritage AI, Egba Adire