When Endangered Voices Speak: Building the First Ehugbo Dialect Audio Dataset Through Grassroots Collaboration

Abstract:
The rapid advancement of speech technologies has created a growing linguistic divide, leaving behind communities whose languages are underrepresented in digital resources. This is acutely true for endangered languages, which often lack a standardized writing system, making them invisible to most natural language processing (NLP) pipelines. We present a case study in grassroots, culturally sensitive dataset creation for the Ehugbo dialect of the Igbo language, spoken by approximately 150,000 people in Afikpo, Ebonyi State, Nigeria. With no prior publicly available audio data, Ehugbo is at high risk of digital extinction.

Our work is motivated by the principle that AI for good must be inclusive and cannot be built solely on the world's majority languages. We describe the collaborative construction of the first open-source Ehugbo audio dataset, a 42.99-minute corpus of biblical recitations. This text was chosen due to the recent (2020) publication of the Ehugbo New Testament, one of the 4 available written resources. The project was built from the ground up by and with the local community: we engaged six native speakers for recording, three validators for quality assurance, and a local sound engineer, ensuring cultural authenticity and community ownership.

To benchmark the dataset's utility for automatic speech recognition (ASR), we evaluated four pre-existing models trained on related languages. The models, `facebook/wav2vec2-xls-r-300m` (fine-tuned on Common Voice Igbo and FLEURS), `AstralZander/xlsr-finetune-Igbo`, `oyemade/w2v-bert-2.0-igbo-CV17.0`, and `CLEAR-Global/w2v-bert-2.0-igbo_naijavoices_250h`, `deepdml/wav2vec2-large-mms-1b-igbo-mix` were chosen for their diverse training data. Results confirm the critical need for language-specific data. The best-performing model, fine-tuned on the Naija Voices corpus (`CLEAR-Global`), achieved a Word Error Rate (WER) of 0.75 and a Character Error Rate (CER) of 0.26. Performance significantly improved when diacritics were removed (WER: 0.70, CER: 0.22), highlighting a key challenge in processing African languages with rich tonal systems. The other models, trained on broader Igbo data, performed worse (WERs from 0.73 to 0.87), underscoring the dialectal uniqueness of Ehugbo.

This project provides a replicable blueprint for building ethical, grassroots datasets for other endangered languages. It demonstrates that overcoming the AI linguistic divide requires more than just model innovation; it demands community-centered collaboration to create the foundational resources that make technological parity possible. We release this dataset to the public to foster research in low-resource ASR and to take a concrete step toward preserving the voice of the Ehugbo people.Our project demonstrates a practical framework for bridging the gap between cutting-edge technology and endangered linguistic heritage, advocating for an AI future where no voice is left behind.