Pidgin Science Voices: A Community-Driven Speech Corpus for Inclusive STEM Education

Introduction

Scientific knowledge in Nigeria is often restricted to academic English, leaving out millions of speakers of Nigerian Pidgin (~75 million people). Over 38 million Nigerian adults remain functionally illiterate, creating a significant accessibility gap in STEM education. Building on our previous work where we collected and translated English Scientific text to Nigerian pidgin needed to build a Machine Translation system that can accurately translate this low-resource language, we extend the work from written translation to speech. The goal is to build the first large-scale, science-focused Nigerian Pidgin speech corpus, enabling automatic speech recognition (ASR), text-to-speech (TTS), and voice-enabled learning tools that democratize scientific knowledge for underrepresented communities.

Methodology

We adopte a facilitator–voice donor model inspired by community "data-farming" approaches.

- **Dataset:** Began with 2,500 Eng-PidginBioData sentences and expanded with 7,500 teacher-reviewed scientific sentences across biology, ecology, nutrition, and genetics.
- **Recording:** 40 speakers being recruited across four Pidgin-dense Nigerian zones (Lagos, Warri, Onitsha, Abuja). Recordings conducted via Lyngual Labs Recorder (48 kHz, 16-bit WAV) with AWS S3 sync; Zoom H1n recorders as backup.
- **Consent & Ethics:** Facilitators first explain the project in Pidgin, ensuring participants fully understand scope and use. Consent is obtained through both oral recording and a signed form, with the right to withdraw at any stage. All data will be released under CC BY-NC-SA 4.0
- **Validation:** Dual human transcription (facilitator + annotator), automatic SNR filtering, and community listening circles for naturalness and intelligibility.
- **Metadata:** Collect speaker demographics (age buckets, gender, accent, environment) to support fair and robust ML modeling

Results

The Pidgin Science Voices (PSV) corpus will comprise ~20 hours of validated speech paired with transcripts and a glossary of Pidgin scientific terms. The collection of this speech dataset will enable pilot experiments showing:

- **ASR Fine-tuning:** Whisper-small trained on PSV achieving recognition accuracy on code-mixed speech compared to the baseline English-only model.
- **TTS Prototypes:** Long-form recordings from three speakers enabling initial VITS-based synthetic speech to produce intelligible, natural-sounding Pidgin scientific explanations.
- **Educational Tools:** Audio flashcards, interactive quiz bots, and narrated science capsules demonstrating immediate applicability for learners and teachers.

Discussion & Conclusion

PSV addresses the critical absence of scientific Pidgin speech resources and provides a replicable framework for building community-driven speech corpora in low-resource settings. By enabling inclusive ASR/TTS research, PSV advances responsible machine learning for underrepresented languages. Future work includes scaling to other STEM domains, expanding regional speaker diversity, and benchmarking downstream models on comprehension tasks.

Key References

[1] Emezue et al., *The NaijaVoices Dataset: Cultivating Large-Scale, High-Quality, Culturally-Rich Speech Data for African Languages*, *ArXiv* 2025.

[2] Radford et al., *Whisper: Robust Speech Recognition via Large-Scale Weak Supervision*, OpenAI, 2022.