Building Low-Resource Speech Pipelines for Yoruba Using Data-Centric Methods

Kolawole Ladipo

December 15, 2025

Abstract

Large-scale language and speech models are primarily optimized for English and a handful of other high-resource languages, leaving many widely spoken languages underrepresented. Yoruba, a Niger–Congo tonal language, exhibits properties such as lexical tone and complex morphophonemics that current multilingual systems handle poorly. While recent foundation models nominally include Yoruba, their practical performance is inadequate for tasks requiring tonal accuracy or semantic precision. These shortcomings arise more from issues of data quality and representation than from limitations in model architecture.

This work outlines a data-centric speech processing pipeline for Yoruba that emphasizes capture-time annotation, tone-preserving transcription, and structured corpus design over novel model development. Speech data is collected from native speakers in the Ibadan region in Nigeria, corresponding to Standard Yoruba, using controlled, tone-marked prompts. We describe a unified pipeline for Automatic Speech Recognition (ASR), Text-to-Speech (TTS), and machine translation, utilizing transfer learning, pseudo-labeling, and multilingual subword modeling. Specific attention is given to phonotactic features like labial–velar consonants and to evaluation metrics beyond word error rate that account for tone.

This paper details the system design and data methodology; empirical results are reserved for future work. The contribution aims to be a reproducible framework showing how targeted, high-quality data collection can enable effective speech technology for low-resource tonal languages.

1 Introduction

Recent progress in speech and language modeling has been largely driven by pretraining on vast, web-scraped datasets, leading to systems that perform well for a limited set of high-resource languages. This approach, however, systematically underperforms for tonal and morphologically complex languages that lack substantial curated data. Yoruba, with over 40 million speakers in West

Africa and the diaspora, remains inadequately served by these models despite its inclusion in some multilingual benchmarks.

Current ASR systems, including Whisper, deliver passable word-level transcription for Yoruba in ideal settings but consistently fail to capture lexical tone and critical phonemic distinctions. Orthographic tone marks are often absent in training corpora, and available speech data is frequently poorly labeled, code-switched, or misaligned. Consequently, standard metrics like word error rate mask semantically significant errors, resulting in misleading performance assessments.

This work adopts a data-centric strategy for Yoruba speech modeling. We collect speech data directly from native speakers in the Ibadan region of Oyo state, Nigeria, corresponding to the standard Yoruba dialect, using carefully tone-marked prompts within a controlled dialect region, thereby avoiding the challenges of post-hoc tone reconstruction. The proposed pipeline integrates ASR, TTS, and machine translation components, with explicit mechanisms to handle Yoruba-specific phonotactics and dialectal variation. Instead of introducing a new model architecture, we focus on corpus design, annotation strategy, and evaluation methodology. Special attention is given to certain consonants (/gb/, /kp/) that are absent from English and to tone-sensitive evaluation metrics beyond word error rate.

This paper focuses on system design and data methodology; empirical results are planned as future work. The intended contribution is a reproducible framework demonstrating how modest amounts of high-quality speech data can enable robust speech technologies for low-resource tonal languages.

2 Background and Related Work

Speech processing for low-resource languages has traditionally depended on techniques like transfer learning, multilingual pretraining, and data augmentation to address scarce labeled data [1,2]. Recent foundation ASR models trained on vast, weakly supervised web audio achieve broad language coverage but show degraded performance on tonal languages, where acoustic errors can preserve word forms while completely altering meaning [3].

In speech synthesis, contemporary models such as diffusion-based and codec-based architectures (e.g., NaturalSpeech 3 [4], E2 TTS [5]) have reduced data requirements and improved naturalness. These models, however, assume accurate phoneme–prosody alignment and are vulnerable to tonal errors in the training material. Their application to Yoruba specifically remains largely unexplored.

Dialect representation in large language models is usually an implicit outcome of data imbalances rather than a designed feature. These models learn to recognize and generate dialectal variations by absorbing patterns from massive, heterogeneous datasets scraped from the internet, which contain uneven mixes of standardized and regional forms. Consequently, a model's capability to understand or produce a specific dialect emerges statistically, depending on

the volume and quality of examples it encountered during training. Previous research indicates such representations tend to favor standardized varieties and can encode biases against non-standard dialects [6]. Though collecting dialect-specific data and performing targeted fine-tuning offers a path to improvement, these methods are difficult to generalize and apply consistently across broad, interconnected dialect landscapes.

3 Methodology and System Design

3.1 Dialect Representation

We model dialectal variation as a bidirectional graph where nodes represent specific dialects (e.g., Ìbàdàn Standard, Ìjẹ̀bú , Èkìtì, Ẹ̀gbá), while edges connecting them encode shared morphosyntactic and lexical features. During an initial, shared pretraining phase, the model learns the core grammatical structure common to all Yoruba dialects. Following this, we employ a lightweight, adaptive fine-tuning mechanism. Instead of retraining a massive model from scratch for each dialect, this method selectively adjusts the model's parameters to learn the specific phonological and prosodic patterns (i.e. characteristic tonal contours or vowel realization) of a target dialect. This graph-based, adaptive approach is far more efficient than traditional methods that rely on parallel corpora (identical content recorded in multiple dialects), which are prohibitively expensive and time-consuming to create at scale. By learning the relationships between dialects, the system can also make more informed predictions for under-resourced varieties, leveraging knowledge from their better-represented neighbors.

3.2 ASR Pipeline

The ASR pipeline uses Whisper as a foundation. However, we make several critical modifications to overcome its limitations with tonal languages like Yoruba. First, we implement a tone-aware text normalization layer that consistently maps various orthographic representations of tone (or the lack thereof) into a standard format. Second, we assemble a curated training dataset of approximately 50 hours of clean, studio-quality Yoruba speech, where every recorded entry is paired with a meticulously tone-marked and phonetically verified transcript. We employ a curriculum fine-tuning strategy, where the model is first exposed to simpler, clearly enunciated speech before gradually introducing more natural conversational variations. Our central hypothesis is that a smaller amount of high-quality, tone-perfect data is actually more useful for learning than vast amounts of messy, unlabeled audio from the web. To give the model even more precise guidance, we use forced alignment techniques to break down the audio into individual sounds and tones, so the model learns not just which word was said, but exactly how it was pronounced. Finally, we advocate for and will implement a suite of evaluation metrics that move beyond the flawed

standard of Word Error Rate (WER), incorporating Tone Error Rate (TER) and minimal-pair accuracy to truly assess semantic understanding.

3.3 TTS Pipeline

TTS system for Yoruba requires a different focus. While these models can learn from fewer hours of audio than speech recognition systems, they demand extremely high-quality recordings to sound natural. We record all our speakers in a sound-treated studio using professional equipment, capturing audio at 24-bit and 48 kHz for maximum clarity. We then downsample this audio as needed for different model requirements. The heart of our effort is a new corpus we're building called **ÌròyìnSpeech**. Its key feature is that for every audio clip, we provide multiple transcript files—one for the standard orthography and others that capture dialect-specific pronunciations. For instance, a single recording might have a `yovo_3501.txt` file for Standard Yoruba and an `ekiti_3501.txt` file showing how the same sentence is spoken in the Ekiti dialect. By training the TTS model on these parallel transcripts during its initial learning phase, we teach it to associate the correct sounds and intonations with each dialect label. This ensures that when we ask the system to generate speech in a particular dialect, it sounds authentic and stays consistent with how our other components, like the translator, understand that dialect.

4 Preliminary Analysis and Expected Outcomes

Pilot experiments are currently planned, not yet completed. Evaluation will utilize a held-out set of phrases not seen during training. Speech generated by the TTS system will be transcribed by a separately trained ASR model, and these transcriptions will be compared against tone-marked reference texts. Planned metrics include Tone Error Rate, minimal-pair accuracy, phoneme confusion rates for critical sounds like labial–velar stops, and human intelligibility scores.

A closed-loop assessment, termed *tone-consistency validation*, as shown in Figure 1, will measure whether synthesized speech retains correct lexical tones when decoded by an ASR system. While WER will be reported, it will be considered a secondary metric. The anticipated outcome is that explicit tone-aware supervision will lead to significant improvements in semantic correctness, even if the surface-level WER shows only minor changes.

5 Discussion and Future Work

Future directions include expanding the corpus to cover additional Yoruba dialects, incorporating speaker adaptation techniques, and assessing the pipeline's utility in downstream applications like language education. Open research questions pertain to the efficiency of cross-dialect knowledge transfer and how tonal supervision interacts with existing multilingual pretraining paradigms. The

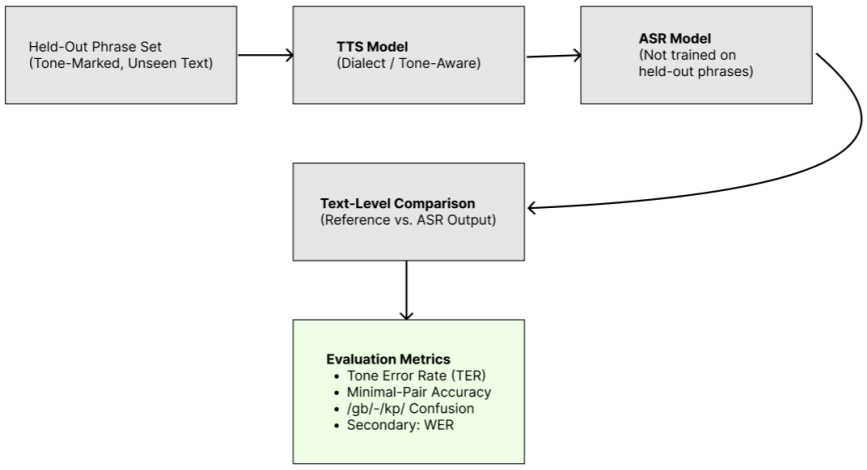

Figure 1: Closed-Loop Tone Consistency Validation

methodological framework described here may be applicable to other Niger–Congo languages with similar tonal properties.

6 Conclusion

This paper proposes a data-centric framework for developing speech technology pipelines for Yoruba. It shifts the focus from architectural innovation to the fundamentals of annotation strategy, corpus design, and meaningful evaluation. By prioritizing tonal fidelity and controlled data collection from the outset, the approach seeks to establish a foundation for robust ASR and TTS performance for a low-resource tonal language. Empirical validation of this framework is designated as future work.

References

[1] Alec Radford et al. Robust speech recognition via large-scale weak supervision. In *Proceedings of ICML*, 2023.

[2] David Ifeoluwa Adelani et al. Masakhane: Machine translation for Africa. In *Proceedings of MT Summit*, 2021.

[3] A. Nguyen et al. Data-centric AI: A survey. *arXiv preprint arXiv:2201.00450*, 2022.

[4] Alexis Conneau et al. Unsupervised cross-lingual representation learning for speech recognition. In *Proceedings of NeurIPS*, 2020.

[5] Shawn Hershey et al. Towards universal speech recognition. *arXiv preprint arXiv:2303.01037*, 2023.

[6] Zeyu Ju et al. NaturalSpeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *arXiv preprint arXiv:2403.03100*, 2024.

[7] Chengyi Wang et al. E2 TTS: Embarrassingly easy fully non-autoregressive zero-shot TTS. *arXiv preprint arXiv:2406.18009*, 2024.

[8] Allison Koenecke et al. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 2020.