Tone in Yoruba ASR: Evaluating the Impact of Tone Recognition on Transformer-Based ASR Models

Joy Olusanya
Department of Linguistics and African Languages
Obafemi Awolowo University
Ile-Ife, Nigeria
joynaomiolusanya@gmail.com

Abstract

This research investigates the role of tone in Standard Yoruba Automatic Speech Recognition (ASR), focusing on how explicit tone marking (diacritics) influences accuracy and overall system performance. As a low-resource tonal language, Yoruba encodes critical lexical and grammatical contrasts via pitch, making tone handling both essential and challenging for ASR. Three pre-trained models, Meta's *MMS-1B-all*, OpenAI's *Whisper-small*, and *AstralZander/Yoruba_ASR*, were trained and evaluated on datasets that vary by tone annotation (fully tone-marked vs. non-tone-marked). Using Word Error Rate (WER) and Tone Error Rate (TER) as primary metrics, results consistently favoured non-tone-marked data, yielding substantially lower error rates than their tone-marked counterparts. These outcomes suggest that current architectures encounter difficulties with diacritically marked Yoruba, likely stemming from tokenisation behaviour, insufficient representation of tonal cues, and limited tone modelling in the underlying pre-training. The study concludes that tone-aware approaches, spanning tokenisation, acoustic-text alignment, and model objectives, are necessary to improve recognition for Yoruba and other low-resource tonal languages. The findings clarify the interaction between linguistic tone systems and computational modelling, and offer concrete directions for building more robust, tone-sensitive ASR systems.

1 Introduction

Automatic Speech Recognition (ASR), also known as voice recognition, is a transformative technology that plays a crucial role in enabling human-computer interaction by converting spoken language into written text. As an independent, machine-based process of decoding and transcribing oral speech (Levis and Suvorov, 2012), ASR has found applications across diverse languages and domains. Its significance extends beyond convenience, as it also contributes to the preservation and documentation of linguistic diversity, particularly for underrepresented languages.

Yoruba, a West Benue-Congo language, is predominantly spoken in the Southwestern region of Nigeria and in several other countries, including Ghana, Togo, the Republic of Benin, Haiti, and Sierra Leone (Fábùnmi, 2013). As a tonal language with a three-tone system, Standard Yoruba (SY) stands to benefit immensely from advances in ASR. Tonal distinctions in Yoruba, high (H), mid (M), and low (L), are crucial in distinguishing meaning between lexical items. The language also exhibits rising and falling tones, adding further complexity. These tonal features are vital not only for everyday communication but also for understanding broader linguistic domains such as phonology, syntax, morphology, semantics, and pragmatics.

Despite the growing interest in African language technologies, a notable gap exists in Yoruba ASR research, especially concerning the treatment of tone marking. Tone marking refers to the explicit representation of tonal distinctions using diacritics in written text. This practice has the potential to enhance clarity and reduce lexical ambiguity in Yoruba. However, the extent to which tone marking influences ASR accuracy and performance has not been sufficiently explored. Addressing this issue is essential to advancing Yoruba ASR and ensuring that technological solutions adequately reflect the linguistic realities of the language.

ASR systems can also contribute to building extensive digital resources for African languages. By capturing tones, pronunciation, and other linguistic nuances, these systems can provide rich datasets that aid both language preservation and technological innovation. For Yoruba, such efforts not only enhance its digital representation but also help ensure that the language remains relevant in

Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026), pages 149–156

the digital age.

This study seeks to bridge the gap by examining how tone and tone marking affect Yoruba ASR models. Specifically, it evaluates whether incorporating tone marking can improve transcription quality and explores the degree to which different levels of tone marking influence system outputs. The insights gained will not only enhance Yoruba ASR but also provide guidance for developing ASR systems for other tonal languages, reinforcing inclusivity and linguistic diversity in speech technologies.

1.1 Aims and Objectives of the Study

The overall aim of this research is to examine the role of tone and tone marking in improving the performance of Yoruba ASR systems. The specific objectives are:

1. To identify the role and significance of tone in Standard Yoruba ASR.
2. To examine the impact of tone marking on ASR models' transcription accuracy.
3. To evaluate the effects of different levels of tone marking in Yoruba ASR datasets on overall system performance.

2 Related Work

Tone plays an important role in Yoruba Automatic Speech Recognition (ASR), as extensively discussed in both linguistic and computational literature. (Van Niekerk and Barnard, 2012) highlighted that while non-tonal languages such as English rely on stress patterns, tonal languages like Yoruba use pitch variations to distinguish lexical meaning. This renders tone not merely a prosodic feature but a lexical and grammatical cue, directly impacting the intelligibility and accuracy of ASR outputs. (Odéjobí, 2008) reinforces this distinction by comparing stress-based contrasts in English (e.g., *record* [noun] vs. *record* [verb]) to tonal contrasts in Yoruba (e.g., *bá* [H: to catch up with], *ba* [M: to roost], and *bà* [L: to perch]). These minimal pairs show that tone alone can differentiate word meaning, making tonal modelling essential for accurate transcription and semantic interpretation in Yoruba ASR systems.

In computational speech research, tone recognition has largely followed data-driven approaches. (Odéjobí, 2008) explored the use of Artificial Neural Networks (ANNs), particularly Multilayer Perceptrons (MLPs) and Recurrent Neural Networks (RNNs), for classifying isolated Yoruba tones based on the fundamental frequency (f0) contour. Both models demonstrated promising accuracy, with RNNs slightly outperforming MLPs due to their strength in modelling sequential data. Unlike traditional Hidden Markov Models (HMMs), which assume conditional independence of features, neural networks are capable of learning non-linear acoustic patterns and modelling coarticulatory and prosodic variations, making them well-suited for tone classification.

However, tone modelling becomes significantly more complex in continuous speech, where factors like tone coarticulation, sandhi, and overlapping intonation contours introduce additional variability. (Ogunremi et al., 2024) addressed this challenge by fine-tuning a pre-trained wav2vec 2.0 model on a Yoruba speech corpus. Their results showed that self-supervised learning of contextual acoustic representations significantly outperformed end-to-end models such as Conformer in terms of ASR accuracy. This supports the hypothesis that incorporating contextual embeddings from pre-trained models can improve tone-sensitive recognition in continuous speech. Nevertheless, the study did not explicitly investigate the effects of tone marking in text transcripts or analyze how tonal distinctions are preserved or lost in model outputs.

Despite these advancements, little effective research has been done on integrating or evaluating tone representations into transformer-based ASR architectures or large language models (LLMs). State-of-the-art models such as wav2vec 2.0 (Baevski et al., 2020) and Whisper (Radford et al., 2023) , though highly effective in general ASR tasks, are typically trained on large, multilingual corpora that either rely on word error rate as the primary metric and omit tone markings, or include poorly annotated tonal data for under-resourced tone languages. As a result, these models are largely tone-agnostic, which limits their performance on languages like Yoruba where tone carries semantic weight.

Recent efforts to refine ASR outputs using LLMs also fall short in addressing tone. For example, (Karner et al., 2024)investigated how LLMs such as SauerkrautLM can post-process ASR hypotheses to correct syntactic and lexical errors in spontaneous speech. While their results showed improvements in semantic similarity metrics, their

framework was based on a German corpus and did not consider tone-sensitive features, nor did it evaluate the implications of tone errors on ASR output meaning an omission that would be critical in tonal contexts.

The literature consistently highlights the linguistic necessity of tone modeling in Yoruba ASR and recognizes the technical potential of neural and transformer-based architectures. Yet, a critical gap persists: transformer-based models and LLM-enhanced ASR systems largely fail to encode or utilize tone as an essential feature, either in their input representations or in evaluation metrics. This research therefore aims to investigate how transformer-based ASR models recognize and handle tone in Yoruba, with a particular focus on the effect of tone marking in Yoruba text on ASR model performance. By explicitly examining whether encoding tonal information improves recognition accuracy and semantic fidelity in continuous Yoruba speech, this study seeks to fill the existing gap and contribute to the development of tone-aware ASR systems for tonal languages.

3 Overview of Yoruba

Standard Yorùbá (SY), a major Niger-Congo language spoken by over 40 million people in Nigeria, Togo, Benin, and across the diaspora, is a tone language with 18 consonants, 7 oral vowels, 5 nasal vowels, and two syllabic nasals. Its orthography consists of 25 letters, including the digraph *gb*, and uses three contrastive tones (High, Mid, Low) alongside Rising and Falling allotones. Tone plays a crucial role in distinguishing both lexical meaning and grammatical functions.

The orthography reflects this phonological complexity through tonal and phonemic diacritics, which are essential for disambiguating words. For example, the string *gba* may mean "receive," "spread," or "forgive," depending on tonal marking. Without diacritics, Yorùbá texts become ambiguous, creating challenges for both human comprehension and natural language technologies such as ASR and machine translation.

From a computational perspective, the lack of diacritics in digital texts poses a major barrier. (Orife et al., 2020) highlights Automatic Diacritic Restoration (ADR) as a critical step toward accurate Yorùbá NLP. Their Transformer-based models and FastText embeddings demonstrate the effectiveness of ADR for improving downstream tasks. They emphasize the importance of diverse training corpora, including literature, blogs, interviews, and proverbs.

Dialectal studies further reveal tonal variation within Yorùbá. (Adeniyi, 2018), focusing on the Ìgbòmina dialect, identifies a process called High Tone Lowering (HTL), where a high tone is lowered to mid-level under specific phonological conditions. This phenomenon influences morphosyntactic structures, particularly in tense marking, and illustrates the deep interaction between tone, phonology, and syntax.

Together, these studies underscore the central role of tone in Yorùbá. For both theoretical linguistics and computational applications, tone and diacritics are indispensable for capturing the richness of the language and for building robust, tone-aware language technologies.

Use of Diacritics to Identify Tone in Yorùbá

Pitch is a central feature in Yorùbá, where it distinguishes lexical and grammatical meaning. The language employs three phonemic tone levels: High (H), Mid (M), and Low (L) (Raji, 2015). These are marked using diacritics in written form, primarily on vowels and syllabic nasals. Accurate usage is essential for conveying intended meanings, particularly where tonal ambiguity could cause semantic confusion.

igba can mean "calabash" (igbá), "climber's belt" (igbà), or "two hundred" (igba), depending entirely on tonal marking (Raji, 2015)

Table 1: Use of Diacritics to Identify Tone in Yorùbá

Diacritic	Description	Word	Gloss
´	High tone	*igbá*	Calabash
`	Low tone	*igbà*	Climber's belt
	Mid tone	*igba*	Two hundred

These tonal markings are indispensable in preventing ambiguity in meaning and preserving the linguistic integrity of Yorùbá. As a register tone language, Yorùbá relies on three pitch levels (High, Mid, Low) to distinguish lexical and grammatical meanings. Unlike stress-based languages, where emphasis lies on syllabic prominence, Yorùbá uses pitch as a core semantic marker (Akinlabi, 2004; Connell and Ladd, 1990) Minimal pairs such as *Sán* ("to cut grass"), *San* ("to bite"), and *Sàn* ("watery") illustrate how tone alone can differentiate

semantically unrelated words with identical segmental makeup. Similarly, the word *igba* may denote "calabash," "two hundred," or "rope for climbing palm trees" depending on tonal assignment (Adeniyi, 2021).

Beyond lexical tone, Yorùbá exhibits tonal processes such as assimilation, downdrift, downstep, and tone raising, which influence connected speech (Laniran and Clements, 2003). Downdrift occurs when successive low tones depress following tones across an utterance (e.g., *àgbàdo* "corn"), while downstep involves the deletion of a low tone (often due to elision) that still influences the pitch of subsequent tones, even though it is no longer phonetically present. This makes downstep a floating tone phenomenon, one that exists abstractly in the tonal. High tone raising is another key phenomenon, wherein a high tone preceding a low tone is raised slightly, likely as a compensatory prosodic adjustment to maintain pitch contrast, such as the word *àgbúngbû*. Similarly, processes like consonant deletion in *Èdúdú* → *èédú* ("charcoal"), further reflecting the intricate interaction between segmental reduction and tonal retention in fluent speech (Adeniyi, 2021).

(Elugbe, 1985) introduces the concept of contour overlap, noting that Yorùbá rising and falling contours are phonetic transitions between level tones rather than independent phonemes. This aligns with (Welmers, 1973), who argued that African tonal languages rely primarily on level tones, even when surface contours are present.

Taken together, the Yorùbá tone system is not an additive prosodic feature but a central grammatical component. Its integration with morphology and syntax makes accurate tonal representation vital for both linguistic description and computational applications, such as Automatic Speech Recognition (ASR), diacritic restoration, and text-to-speech technologies.

4 Dataset

The data for this study was collected from the Common Voice Yoruba dataset platform, a publicly available voice dataset powered by contributions of volunteers across the globe. Common Voice aims to democratize voice technology by offering open-access speech data for training machine learning models, particularly for low-resource languages and communities. The dataset consists of 3,099 utterances, corresponding to approximately 6 hours of validated Yoruba speech recordings. This dataset comprises high-quality Yoruba speech recordings, along with corresponding transcriptions, making it suitable for training and evaluating Automatic Speech Recognition (ASR) models. To analyse the role of tone in Yoruba ASR, the transcription from this dataset was manually processed to create different versions based on tone marking. The original transcription, including tone markings, was retained and analysed as a single version. Additionally, a manual process was undertaken to remove tone markings, creating a version of the dataset that represents Yoruba text without tone information. This approach allows for a controlled comparison of ASR performance with and without tone marking. The speech data were carefully segmented and labelled to ensure that each recorded utterance was correctly mapped to its corresponding transcription. Quality control measures were implemented to verify the accuracy of the modified transcription by cross-referencing with Standard Yoruba linguistic resources. The recordings were also examined for clarity, ensuring that they contained minimal background noise and accurately reflected Yoruba phonetic and tonal distinctions. In parallel with sentence-level transcriptions, the word-level dataset was also manually extracted and annotated. Each word was carefully marked for tonal patterns and aligned with its audio counterpart to support evaluation, enabling the model to attempt recognition without prior fine-tuning specifically on tonal forms. By using these datasets and manually modifying the transcription, this study establishes a structured approach to evaluating the impact of tone marking on Yoruba ASR performance, providing insights into the role of tone in ASR accuracy.

5 Methods

The method of data analysis in this study systematically evaluates the role of tone in Yoruba Automatic Speech Recognition (ASR) by comparing how different levels of tone marking influence model performance. The process begins with preprocessing, where Yoruba speech recordings and transcriptions are cleaned, normalised, and prepared for training. Audio signals are converted into machine-readable features using Mel-Frequency Cepstral Coefficients (MFCCs), while transcriptions are tokenised and formatted to ensure tone consistency. Datasets include both tone-

marked and non-tone-marked versions for comparative training.

Three ASR models were examined: Meta's facebook/mms-1b-all, OpenAI's Whisper, and the Yoruba-specific AstralZander/yoruba_ASR. MMS-1b-all extends the wav2vec 2.0 architecture with adapter modules, trained on more than 500,000 hours of speech across 1,400 languages. It emphasises multilingual adaptability with efficient fine-tuning. Whisper, in contrast, is a transformer-based encoder-decoder model trained on 680,000 hours of multilingual, multitask labelled audio. It offers robustness to noisy input but lacks modularity for language-specific fine-tuning. The AstralZander/yoruba_ASR model diverges by focusing exclusively on Yoruba. Fine-tuned from wav2vec2-xls-r-300m, it captures phonetic and tonal subtleties despite relying on smaller training data. Together, these models highlight trade-offs between multilingual generalisation, large-scale weak supervision, and domain-specific adaptation.

During preprocessing, audio was resampled to 16 kHz to standardise inputs, with noise reduction applied to improve clarity. To capture tonal features, a tone mapping dictionary (TONE_MAP) linked Yoruba vowels to their tonal categories: High, Low, or Mid. A function, extract_tone_sequence, generated tonal sequences from texts, which were evaluated against references using Levenshtein distance. This quantified tonal accuracy by measuring the minimal edits required to match predicted tone sequences with true patterns. Data was split into 80% training and 20% validation.

Training employed the Wav2Vec2 model with Connectionist Temporal Classification (CTC) loss. Fine-tuning was conducted on Yoruba data with a batch size of four, three epochs, and a learning rate of 1e-5. Training ran on Google Colab Pro with GPU acceleration, costing approximately $50. Weights & Biases (W&B) was used for experiment tracking and visualisation.

Model evaluation employed both Word Error Rate (WER) and Tone Error Rate (TER). WER measured insertions, deletions, and substitutions relative to the reference text, while TER quantified tonal misclassifications, reflecting the system's ability to recognise pitch contrasts crucial to Yoruba meaning. Further analysis involved confusion matrices to reveal common tone errors and phoneme-level checks for tonal assignment accuracy. Comparative experiments assessed whether tone marking improved ASR accuracy, with hypothesis testing applied to validate significance. Visualisation tools, such as accuracy graphs and error rate plots, supported interpretation. Manual evaluation complemented automated metrics to ensure practical usability.

Through this methodology, the study investigates whether explicit tone marking significantly enhances Yoruba ASR and identifies how different modelling strategies handle tonal complexity.

6 Results

6.1 Overview of Model Performance

This section presents the Word Error Rate (WER) and Tone Error Rate (TER) for three ASR models: Whisper-small, MMS-1B-all, and AstralZander/yoruba_ASR. Results are reported separately for the tone-marked and non-tone-marked Yoruba speech datasets to allow for a clear comparison of performance across different linguistic input conditions. Tables 2 and 3 summarize the results.

Table 2: Performance on Tone-Marked Yoruba Dataset

Model	WER (%)	TER (%)
Whisper-small	65.70	32.31
MMS-1B-all	85.08	52.37
AstralZander/yoruba_ASR	76.90	30.94

Table 3: Performance on Non-Tone-Marked Yoruba Dataset

Model	WER (%)	TER (%)
Whisper-small	37.90	4.20
MMS-1B-all	50.27	5.20
AstralZander/yoruba_ASR	47.52	6.44

Tables 2 and 3 present performance metrics for models on tone-marked and non-tone-marked Yoruba ASR datasets.

6.2 Word Error Rate (WER) Analysis

All three models show consistently higher word error rates on tone-marked datasets compared to the non-tone-marked datasets. For the AstralZander/yoruba_ASR model, the WER increases from 47.52% without tone marking to 76.90% with tone marking. The Whisper-small model also exhibits a similar pattern, with WER rising from 37.90% without tone to 65.70% with tone marking.

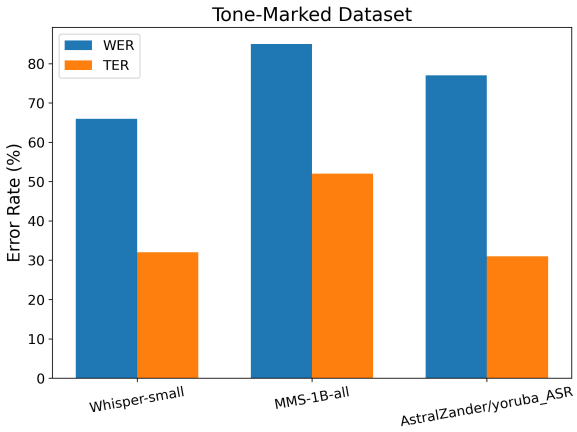

Figure 1: ASR performance on tone-marked Yoruba speech.

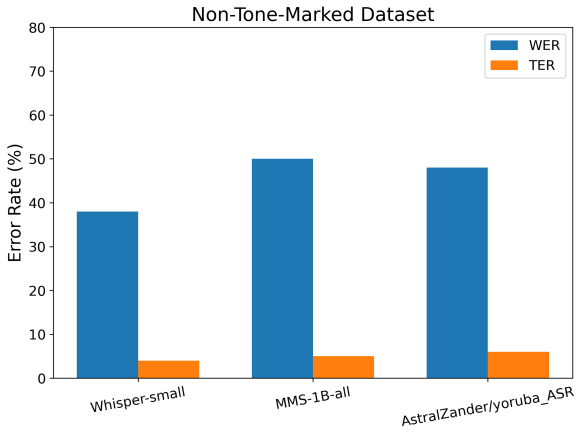

Figure 2: ASR performance on non-tone-marked Yoruba speech.

The MMS-1B-all model experiences the largest increase, with WER increasing from 50.27% without tone to 85.08% with tone marking. These results suggest that the inclusion of tone markings introduces additional complexity in speech recognition tasks, which the current models struggle to handle effectively, leading to poorer word recognition accuracy.

6.3 Tone Error Rate (TER) Analysis

The tone error rate demonstrates a marked increase in tone-marked datasets relative to non-tone-marked datasets across all three models. Specifically, the AstralZander/yoruba_ASR model's TER rises from 6.44% without tone to 30.94% with tone marking. Similarly, the Whisper-small model's TER increases from 4.20% without tone to 32.31% with tone marking. The MMS-1B-all model shows the most significant jump, with TER increasing from 5.20% without tone to 52.37% with tone marking. This significant increase indicates that while tone is an important linguistic feature in Yoruba, the models currently do not effectively capture or utilise tonal information. The rise in tone error rates highlights the challenges these ASR architectures face when processing tone-marked speech data.

6.4 Discussion

6.4.1 RQ1: What is the place of tone in Yoruba ASR models?

Tone is central to Yoruba, serving as a primary cue for distinguishing lexical meaning. However, our findings show that current ASR models (Whisper-small, MMS-1B-all, AstralZander/yoruba_ASR) fail to effectively utilise tonal information. Instead, tone-marked input leads to significantly higher WER and TER, suggesting that tone is treated as noise rather than as a meaningful phonemic feature. This aligns with prior work (Osakuade and King, 2024), which reported that discretised speech representations often lose tone distinctions, even in Yoruba-specialised systems.

6.4.2 RQ2: How does the inclusion of tone marking affect the output and accuracy of Yoruba ASR models?

Across all models, the inclusion of tone markings consistently degraded performance. WER increased by more than 20% for each model, while TER rose sharply (e.g., MMS-1B-all: 5.20% $\rightarrow$ 52.37%). These results highlight that although tone is linguistically necessary for preserving meaning, current ASR architectures cannot yet capture its acoustic and orthographic complexities. This confirms earlier observations Imam et al., 2025, that tonal languages require specialised modelling strategies. Thus, the findings emphasise the need for tone-aware ASR approaches.

6.4.3 RQ3: What are the effects of different levels of tone marking on Yoruba ASR performance?

To analyse the effect of tone marking, a subset of Yorùbá word-level utterances with varying degrees of tonal specification was evaluated using two transformer-based ASR models: *AstralZander* and *MMS-1B-all*. Model outputs were compared against ground-truth references using similarity scores, which measure both lexical accuracy and tonal preservation.

Across all tone conditions, AstralZander consis-

tently outperformed MMS-1B-all, particularly in maintaining tonal and prosodic structure. MMS-1B-all frequently produced tone-neutral or incorrect forms, leading to semantic drift and homophone collisions. For example, the tone-marked word *'Òkú'* ("corpse") was rendered as *àkù* by MMS-1B-all, collapsing tonal distinctions and introducing ambiguity with *àkù*. Similarly, *ògùn* ("medicine" or "herbal remedy") was transcribed as *ogun*, which could also mean "war" or the deity "Ogun," illustrating how tonal errors create semantic ambiguity. Phonological simplification was also observed, such as the deletion of the initial vowel in *Ọ̀tẹ̀* ("rebellious"). These examples demonstrate that ASR systems that fail to capture tonal information can produce outputs that are phonologically plausible but semantically incorrect, highlighting the critical role of tone in preserving meaning in Yoruba. Overall, higher levels of tone marking reveal substantial weaknesses in current ASR architectures. MMS-1B-all shows marked degradation as tonal complexity increases, whereas AstralZander demonstrates greater robustness, producing transcriptions with higher tonal fidelity and fewer tone-induced semantic errors.

6.4.4 Summary of Findings

Overall, tone markings substantially reduce ASR performance across all models tested. Whisper-small, MMS-1B-all, and AstralZander/yoruba_ASR demonstrate limited tone sensitivity, treating tonal cues as noise rather than as essential linguistic information. These results reinforce the urgent need for tone-aware ASR models capable of addressing the acoustic and linguistic challenges inherent in tonal languages such as Yoruba.

7 Conclusion

This study examined the impact of tone and tone marking on Yoruba ASR performance. Results indicate that transformer-based ASR systems generally perform better with non-tone-marked input, as the inclusion of diacritic tone marks often increases error rates. While some models, particularly AstralZander/yoruba_ASR, show a relatively better ability to preserve tonal and prosodic patterns, issues such as homophone collisions, tonal flattening, and phonological simplifications remain frequent.

While current ASR systems demonstrate promising capabilities for Yoruba speech transcription, they lack sufficient sensitivity to tonal features. Effective Yoruba ASR must go beyond phoneme recognition and integrate tonal modeling as a core component of its architecture. Future development should include tone-aware training data (both tone-marked and non-tone-marked), specialized tokenizers for tonal languages, open access to well-annotated Yoruba speech corpora, and evaluation metrics beyond traditional Word Error Rate (WER), such as Tone Error Rate (TER) or Semantic Error Rate (SER). Interdisciplinary collaboration among linguists, AI researchers, and software developers is crucial to ensure the linguistic complexity of Yoruba is adequately captured. Additionally, exploring transformer-based models built from scratch may help optimize ASR systems for tone-sensitive tasks and improve generalization in low-resource settings.

Limitations

This study is limited by the scope of available datasets, which may not capture the full range of Yoruba dialectal and contextual variation. In particular, no dialectal variation was included in this work, and future research may need to incorporate multiple Yoruba dialects to improve generalisation. Moreover, the pre-trained models evaluated were not fully originally optimised for tonal representation, potentially biasing performance against tone-marked data. Finally, the evaluation metrics (WER and TER) provide only a partial assessment of model performance, suggesting that future work could further explore tone-aware architectures.

Acknowledgments

I would like to thank my supervisor, Dr Kolawole Adeniyi, for his guidance, detailed feedback, and insightful discussions throughout this project.

References

K. Adeniyi. 2018. High tone lowering in Ìgbòmina-yorùbá. *Journal of West African Languages*, 45(2).

K. Adeniyi. 2021. A tonal identification of yoruba dialects. *Dialectologia*, 27:1–31.

A. Akinlabi. 2004. The sound system of yoruba. In N. S. Lawal, M. N. O. Sadiku, and P. A. Dopamu, editors, *Understanding Yoruba Life and Culture*. Africa World Press.

Alexei Baevski, Hao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A frame-

work for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

B. Connell and R. D. Ladd. 1990. Aspects of pitch realization in yoruba. *Phonology*, 7(1):1–29.

B. Elugbe. 1985. The tone system of ghotuo. *Cambridge Papers in Phonetics and Experimental Linguistics*, 4(1):1–21.

F. A. Fábùnmi. 2013. Negation in sixteen yorùbá dialects. *Open Journal of Modern Linguistics*, 3(1):1–15.

F. Imam, I. Nurideen, I. Orife, and B. F. Dossou. 2025. Automatic speech recognition for african low-resource languages: Challenges and future directions.

M. Karner, S. Müller, and C. Klein. 2024. Towards improving asr outputs of spontaneous speech with llms. In *Proceedings of the Annual Conference on Speech and Language Technology*.

Y. O. Laniran and G. N. Clements. 2003. Downstep and high raising: Interacting factors in yoruba tone production. *Journal of Phonetics*, 31:203–250.

J. Levis and R. Suvorov. 2012. Automatic speech recognition. In *The Encyclopedia of Applied Linguistics*. Wiley.

À. Ọ. Odéjobí. 2008. *Recognition of tones in Yorùbá speech: Experiments with artificial neural networks*. Ph.D. thesis.

T. O. Ogunremi, K. Tubosun, A. Aremu, I. Orife, and D. I. Adelani. 2024. Ìròyìnspeech: A multi-purpose yorùbá speech corpus. In *Proceedings of LREC-COLING*.

Iroro F. Orife, David I. Adelani, Timi Fasubaa, Victor Williamson, Wuraola Fisayo Oyewusi, Olamilekan Wahab, and Kola Tubosun. 2020. Improving yorùbá diacritic restoration. *arXiv preprint arXiv:2003.10564v1 [cs.CL]*.

O. Osakuade and S. King. 2024. Do discrete self-supervised representations of speech capture tone distinctions?

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2023. Whisper: General-purpose speech recognition. In *Proceedings of ICML*.

H. O. Raji. 2015. *The tone system of Arogbo-Ijaw*. Ph.D. thesis.

D. R. Van Niekerk and E. Barnard. 2012. Tone realisation in a yoruba speech recognition corpus. In *Proceedings of the Third International Workshop on Spoken Language Technologies for Under-resourced Languages*, pages 54–59.

W. E. Welmers. 1973. *African language structures*. University of California Press.