ÒWE-Voice: An Evaluation of Monolingual and Multilingual ASR Model Using Yoruba Proverb Speech Dataset

Daud Abolade
Masakhane
NITHUB, University of Lagos
NKANDA
aboladedawud@gmail.com

Abstract

Given the advancement of various Artificial Intelligence (AI) technologies in the 21st century, Automatic Speech Recognition (ASR) plays a vital role in human and machine interaction and serves as an interface for a wide range of applications. The development of these high-performing, robust and useful technologies continue to gain more attention on high-resource languages due to high availability of language data, market profitability dominance and access to funding and research initiatives compared to the marginalised low-resource languages. Despite efforts to develop ASR systems for African languages, there are still numerous challenges due to limited speech datasets, tonal complexity and dialectal variation. In this study, we curated a domain-specific speech dataset for one of the oral Yoruba literatures, proverbs, which are highly culturally inclined. We used the Yoruba recording app that was developed for Iroyin-speech project to record 6 hours of Yoruba proverb sentences. The NCAIR1/Yoruba-ASR model which was finetuned on Open AI Whisper Small and Massively Multilingual Speech, a multilingual speech model featuring low-resource languages including Yoruba language was evaluated with the recorded Yoruba proverbs. Evaluation was conducted based on Word Error Rate (WER) and Tone Error Rate (TER). Our result shows that current ASR systems that support Yoruba does not capture cultural nuances. These findings highlight an urgent need to curate more robust speech datasets that are culturally embedded for low resource languages and in this case particularly, Yoruba language in order to build technological tools that preserve African culture, language and identity.

1 Introduction

Prior to the widespread adoption of literacy in Africa, knowledge systems were transmitted primarily through oral traditions. Among these traditions, proverbs constitute one of the most significant forms of cultural knowledge. Proverbs, alongside with folktales, praise poetry and the Ifa oral corpus is beyond mere creative expressions. It serves as a cognitive tool for teaching, solving problems, reasoning and preserving the language and cultural identity. Orality is African heritage, through societal observations proverbs are born. In Yoruba culture, proverbs hold a special place because they convey values, ideas, logic, expressions, and experiences with figurative language which make it difficult for non-native speakers and native speakers that are not well grounded in the language to decode its meaning (Olusanya et al., 2025). Their meaning often depends on logical reasoning which makes it challenging for both humans and machines.

With the rapid expansion of speech technologies across the globe, the absence of accurate and culturally inclined automatic speech recognition (ASR) systems for African languages presents a major barrier to digital preservation of indigenous knowledge thereby contributing to the digital divide. One of the

major solutions is to create culturally inclined datasets and incorporate them into models for more inclusivity. While there have been series of development in building general purpose speech synthesis and speech recognition system which has resulted to the birth of digital products such as Siri, Alexa, Google's Gemini, and Microsoft Cortana for high-resource languages, African languages still struggle to get to this level due to limited amount of quality speech corpora. It must be mentioned that many African oral traditions are at risk due to rapid urbanization, linguistic shift, and diminishing transfer of indigenous knowledge across generations. Recent advances in multilingual speech models such as Whisper, MMS and SeamlessM4T show promising opportunities for many low resource languages (Radford et al., 2022; Pratap et al., 2023; Communication et al., 2023). However, their performance on culturally nuanced dataset remains unexplored and African oral literature remains challenging for ASR systems because it often deviates from everyday conversational patterns, contains uncommon lexical items and relies heavily on tonal accuracy to preserve semantic meaning. There have been efforts towards the curation of speech dataset to train, evaluate and fine-tune speech synthesis and speech recognition models for low-resource languages, (Junczyk, 2024; Emezue et al., 2025; Oliveira et al., 2023) yet, these datasets often neglect African oral literature content which is the bedrock of Africa's rich culture.

To address this challenge, this study presents Òwe-voice, a 6 hours Yoruba proverb speech datasets as well as an experimentation of two state-of-the-art ASR models, evaluated on a culturally inclined speech dataset of 1,250 recorded Yoruba proverbs. We evaluate NCAIR1/Yoruba-ASR which is a Yoruba specific model and MMS-1b-all using WER and TER. [1]

2 Yoruba language

The Yoruba language is spoken in 10 states in southwestern Nigeria and in some communities in the republics of Benin and Togo with over 40 million native speakers. A language that belongs to the Niger-Congo family, it has about 20 dialects and it is one of the national languages of Nigeria also spoken in other countries like Ghana, Côte d'Ivoire, Sierra Leone, Cuba and Brazil (Owolabi, 2006) which makes it one of the prominent and most widely spoken African languages in the world. The language has 25 letters of the Latin alphabet including additional letters containing subdots, such as (e., gb, ṣ, and o.). Yoruba is a tonal language, meaning that it has three distinctive tone levels-high, mid, and low that are decisive in word distinction. High and Low tones are marked with acute (´) and grave (‘) diacritics respectively, while Mid tone is typically unmarked in standard orthography. Accurate pronunciation depends greatly on the tonal marks and subdots. Yoruba is a culturally rich language which has its own special way of preserving and passing oral knowledge before colonization.

3 Related works

Several research has highlighted the need for a large amount of both textual and speech dataset to build ASR models. Earlier work has focused on creating general purpose speech dataset across the three major Nigerian languages (Igbo, Hausa and Yoruba). (Ogunremi et al., 2024) Created about 42 hours of speech data recorded by 80 volunteers, and 6 hours of validated recordings of news and creative writing domains. (Meyer et al., 2022) A religious domain dataset of 86 hours open speech dataset for ten languages spoken in Sub-Saharan Africa where they trained the VITS end-to-end speech synthesis model. (van Niekerk et al., 2015) this dataset was claimed to be curated for speech recognition research, about 33 diverse speaks both male and female gender. The current largest speech corpus for Hausa, Igbo and Yoruba is (Emezue et al., 2025), where a dataset creation process known as data farming was implemented to curate 1839 hours of speech recording on several domains. (Ahia et al., 2024) introduced a parallel text and speech corpus of standard Yoruba and its dialects to perform a machine translation,

[1] https://github.com/Holuwasege/OWE-Voice-Evaluation

automatic speech recognition, and speech-to-text translation task.

While all these studies have greatly explored curation of speech dataset with different modalities, substantial amounts of speech data to train ASR models for Nigerian languages and other low resource African languages are still limited.

4 ÒWE-Voice Dataset

4.1 Textual Data Preparation

The proverb text used in the development of the Òwe-Voice, Yoruba Proverb Speech Dataset was obtained from two different sources. Our primary textual source was Òwe-Yor, a Yoruba text classification dataset containing both proverbial and non-proverbial sentences (Olusanya et al., 2025). We filtered the dataset to isolate only texts containing proverbs. To extend the coverage of the dataset beyond corpus-based collections, we conducted community fieldwork in Lagos and Ogun State. We engaged both the elders and young native speakers who exhibit strong cultural knowledge and oral tradition competence. Through the short interview session that was conducted, we collected additional 70 proverbs. The fieldwork was adopted to capture proverbs that do not appear in existing proverbial datasets. Manual cleaning and orthographical verification, following the standard Yoruba orthography with the accurate tone marks were conducted on the collected data.

4.2 Speech Data Recording

To create high-quality audio datasets of ÒWE-Voice, we followed the process used for creating IroyinSpeech. Yoruba native speakers who are linguistics students at University of Lagos volunteered and they were tested to confirm and verify their competence in reading Yoruba language. A total of 12 voice talents were engaged, representing both male female speakers to ensure acoustics diversity. The recording was conducted in a controlled environment using an Audio-Technical AT2020USB-X microphone and a quiet studio setup to minimize background noise. The Yorùbá Voice SpeechRecorder that was developed for Iroyin speech (Orife et al., 2022) The app processes a text file, typically containing 250 sentences, by sequentially displaying each line for voice input. Functionality includes: recording, playback, file management and deletion, in the case of multi-take recordings. To standardize the recording protocol we adopted some modality which are:

1) Each speaker received a batch of 250 proverbs.
2) Sentences were displayed with the correct tone mark and its standard orthography.
3) All audio was recorded at 48 kHz, 16-bit WAV format.
4) The voice talent was guided by a prompter who was both a linguist and technically sound in operating the Yorùbá Voice SpeechRecorder.

The final dataset contain 6hrs Yoruba proverb audio samples, each paired with its transcription.

Source	Number of Proverbs
Òwe-Yor Corpus	4,930
Fieldwork Collection	70
Total	5000

Table 1: Dataset curated via Òwe Corpus and Fieldwork Collection.

Figure 1: An illustration showing the environment setup and how Òwe-Voice was curated.

5 Experiment and Result

In this study, we conducted a zero-shot evaluation of the following existing ASR models on Òwe-Voice dataset. We sampled 25% of the utterances across all speakers for evaluation, ensuring that each speaker was represented in the evaluation split. The remaining 75% of the data was reserved for potential fine-tuning

Models	WER	TER
NCAIR1/Yoruba-ASR	72.45	27.83
MMS-1b-all	95.42	66.75

Table 2: Result of the models evaluated on 25% Òwe-Voice dataset

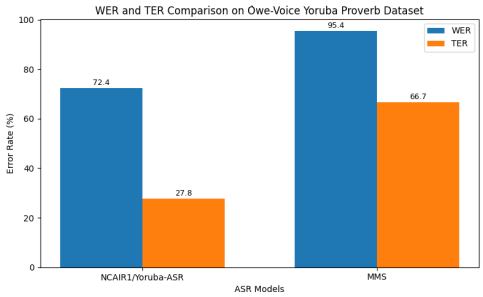

Figure 2: Visualization of the WER and TER score

experiments.

NCAIR1/Yoruba-ASR-v1.0 (Awarri Technologies & National Centre for Artificial Intelligence and Robotics (NCAIR), 2025): This is a monolingual automatic speech recognition (ASR) model finetuned on the Whisper Small architecture, specifically for Yoruba language which is expected to capture linguistics patterns and orthographical conventions.

MMS (Pratap et al., 2023): This is Meta's open-source 1B parameter wav2-vec2 architecture (Baevski et al., 2020) model, supporting 1162 languages, including Yoruba language.

Comparing NCAIR1/Yoruba-ASR with MMS-1b-all allows us to investigate the current state of this model's output on Yoruba proverb speech dataset as part of the indigenous oral knowledge of the Yoruba people.

5.1 Word Error Rate

Word Error Rate (WER) was used as the primary metrics to quantify transcription accuracy at the lexical level. The NCAIR1/Yoruba-ASR model achieved a 72.45% WER which performs better than the multilingual speech model. Although the high error rate shown in this result indicates that the model still struggles with proverbial expressions even though the model was fine-tuned specifically on Yoruba language. On the other hand, MMS produced a much higher WER of 95.42%. This result confirms that multilingual speech models, despite its large training coverage, lack sufficient representation of Yoruba orthography. The model misrecognized common Yoruba lexical items and produced high deletion and substitutions errors. Overall, the WER analysis shows that both models struggle significantly with transcribing Yoruba proverb speech, with the Yoruba specific model offering only partial improvements while the multilingual model largely fails to generalize. These results highlight the difficulty of ASR for low-resource tonal languages.

5.2 Tone Error Rate

Going beyond lexical investigation, we evaluated tonal accuracy using the TER, a metrics that compares predicted tone sequences to reference tone sequences extracted from the predicted transcript of the model and the reference transcript curated alongside with the speech dataset. TER provides insight into the model's ability to preserve Yoruba prosodic distinctions which is a crucial phenomenon because tonal differences often signal different meanings in proverbial expressions. The NCAIR1/Yoruba-ASR model revealed a lower TER score than MMS, showing that the model captures Yoruba tonal patterns better than the MMS model. However, its overall tone accuracy remained challenging. Many tone errors occurred even when the segmental transcription was correct, indicating that the model can predict the right word but fail to assign the correct tone. The MMS model displayed a significantly higher TER, demonstrating very poor tonal generalization. This is expected because multilingual training typically does not emphasize tonal information, and MMS does not explicitly model Yoruba tones. In conclusion, the TER results reveal that tonal errors remain a major bottleneck for Yoruba language in ASR model performance.

6 Beyond Experiments

Yoruba proverb dataset by participating in a Speech Hackathon where we developed a Yoruba Proverb Text-to-Speech web application aimed at evaluating the performance of the Spitch TTS model, a commercial speech technology model built by a language technol-

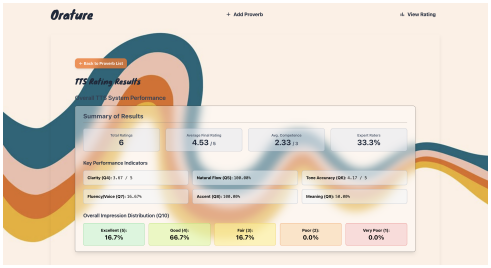

Figure 3

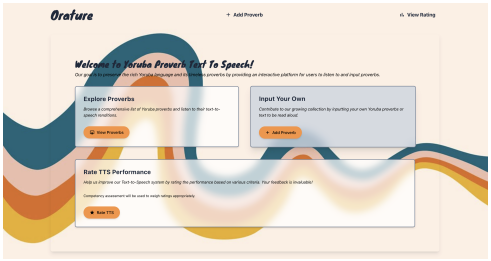

Figure 4: Landing page of Yoruba Proverb Text-to-Speech web application

ogy company in Nigeria. The web app was designed to output spoken Yoruba proverbs directly from the text and to provide an interface through which users could interact with the model.

The primary motivation behind this hackathon was twofold. First, it offered a practical opportunity to assess how well an industry grade TTS model can handle the prosodic complexities of Yoruba language, especially the proverbial expressions, since we are having language tech start ups coming up in Nigeria. Second, it aligned with our mission of preserving Yoruba oral knowledge through technology. [2]

7 Conclusion and Future work

This work introduced Òwe-Voice, a Yoruba proverb speech dataset and the dataset was used to evaluate the performance of both monolingual and multilingual ASR models on culturally rich, low-resource linguistic material. Yoruba proverbs are structurally complex which makes it challenging for ASR models trained on general speech. Due to lack of enough resources and access to GPU we couldn't fine-tune the models. Our evaluation demonstrated that the monolingual NCAIR1/Yoruba-ASR model outperformed the MMS model, having a lower Word Error Rate (WER) and Tone Error Rate (TER) score. Building on this work, We plan to expand the proverb dataset to other African languages and also curate large hours of speech datasets to train existing contemporary ASR systems, including large language model-based speech systems such as GPT-4o/GPT-5-style multimodal models, Qwen3-ASR, Whisper variants, and commercial ASR APIs. We also aim to build other datasets that target African oral knowledge such as praise poetry, folklores and restructure that Yoruba proverb web app. Òwe-Voice can also be explored in other wide range of tasks such as speech translation.

Limitations

There are several limitations in this study. Firstly, due to constraints in computational resource, we were unable to fine-tune the evaluated models before evaluation. Secondly, the scope of the dataset was limited due to the challenges faced in gathering the dataset. Although there was a focus on finding culturally grounded Yoruba proverbs, time, logistics, and budget limitations made it impossible to gather a large amount of speech data. This constraints also prevented the expansion of the dataset to other African languages that have a rich oral literature.

Thus, even though the Òwe-Voice dataset is an important resource that allows one to critically assess ASR models on Yoruba cultural nuances, the corpus is yet to adequately document the African oral knowledge system at large.

References

Orevaoghene Ahia, Anuoluwapo Aremu, Diana Abagyan, Hila Gonen, David Ifeoluwa Adelani, Daud Abolade, Noah A. Smith, and Yulia Tsvetkov. 2024. Voices unheard: Nlp resources and models for yorùbá regional dialects. *Preprint*, arXiv:2406.19564.

Awarri Technologies & National Centre for Artificial Intelligence and Robotics (NCAIR). 2025. Yoruba-asr v1.0: Automatic speech

[2] https://orature.vercel.app/

recognition for yoruba language. Hugging Face Model Hub. Accessed: 2025-12-09.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Preprint*, arXiv:2006.11477.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, and 46 others. 2023. Seamless: Multilingual expressive and streaming speech translation. *Preprint*, arXiv:2312.05187.

Chris Emezue, NaijaVoices Community, Busayo Awobade, Abraham Owodunni, Handel Emezue, Gloria Monica Tobechukwu Emezue, Nefertiti Nneoma Emezue, Sewade Ogun, Bunmi Akinremi, David Ifeoluwa Adelani, and Chris Pal. 2025. The naijavoices dataset: Cultivating large-scale, high-quality, culturally-rich speech data for african languages. *Preprint*, arXiv:2505.20564.

Michał Junczyk. 2024. Framework for curating speech datasets and evaluating asr systems: A case study for polish. *Preprint*, arXiv:2408.00005.

Josh Meyer, David Ifeoluwa Adelani, Edresson Casanova, Alp Öktem, Daniel Whitenack Julian Weber, Salomon Kabongo, Elizabeth Salesky, Iroro Orife, Colin Leong, Perez Ogayo, Chris Emezue, Jonathan Mukiibi, Salomey Osei, Apelete Agbolo, Victor Akinode, Bernard Opoku, Samuel Olanrewaju, Jesujoba Alabi, and Shamsuddeen Muhammad. 2022. Bibletts: a large, high-fidelity, multilingual, and uniquely african speech corpus. *Preprint*, arXiv:2207.03546.

Tolulope Ogunremi, Kola Tubosun, Anuoluwapo Aremu, Iroro Orife, and David Ifeoluwa Adelani. 2024. Ìròyìnspeech: A multi-purpose yorùbá speech corpus. *Preprint*, arXiv:2307.16071.

Frederico S. Oliveira, Edresson Casanova, Arnaldo Cândido Júnior, Anderson S. Soares, and Arlindo R. Galvão Filho. 2023. Cml-tts a multilingual dataset for speech synthesis in low-resource languages. *Preprint*, arXiv:2306.10097.

Joy Olusanya, NLP Masakhane, and Daud Abolade. 2025. Owe-yor: Leveraging transformer based models for yoruba proverb classification. In *Proceedings of the Conference*.

Iroro Orife, Aremu Anuoluwapo, Kọ́lá Túbọ̀sún, , David Ifeoluwa Adelani, and Tolúlọpẹ́ Ógúnrẹ̀mí. 2022. Yorùbá voice speech recorder.

K. Owolabi. 2006. Yoruba. In Keith Brown, editor, *Encyclopedia of Language&Linguistics (Second Edition)*, second edition edition, pages 735–738. Elsevier, Oxford.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling speech technology to 1,000+ languages. *Preprint*, arXiv:2305.13516.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Daniel van Niekerk, Etienne Barnard, Oluwapelumi Giwa, and Azeez Sosimi. 2015. Lagos-NWU yoruba speech corpus. SADiLaR Language Resource Repository, License: Creative Commons Attribution 2.5 South Africa License: http://creativecommons.org/licenses/by/2.5/za/legalcode.