Closing the Gap in Low-Resource ASR: Leveraging Multilingual Models for Code-Switched Yoruba-English Speech

Emmanuel Bolarinwa EMMANUEL@LYNGUALLABS.ORG
Oreoluwa Babatunde OREOLUWA@LYNGUALLABS.ORG
Victor Olufemi VICTOR@LYNGUALLABS.ORG
Kausar Moshood KAUSAR@LYNGUALLABS.ORG
Oluwademilade Williams WILLIAMS@LYNGUALLABS.ORG
LyngualLabs

Abstract

Recent advancements in Automatic Speech Recognition (ASR) have revolutionized voice-based technologies, yet challenges persist in achieving accurate recognition for multilingual and low-resource languages. This research explores the performance of state-of-the-art multilingual ASR models (Whisper Large v3 and MMS-1B-All) on Yoruba-English code-switched (CS) speech. Despite notable progress in multilingual ASR, code-switching remains a complex challenge due to the linguistic intricacies introduced by phonetic, syntactic, and lexical shifts within single utterances. This study addresses a significant gap in the literature by evaluating these models on a 21-hour Yoruba-English dataset and fine-tuned for domain-specific performance. Results show that fine-tuning led to substantial improvements in Word Error Rate (WER), with MMS-1B-All achieving a 55.8% reduction and Whisper Large v3 showing a 50.1% reduction. Although MMS-1B-All outperformed Whisper Large v3 slightly, both models demonstrated strong potential for ASR in Yoruba-English CS speech recognition. This study highlights the feasibility of fine-tuning multilingual ASR models for low-resource code-switched scenarios and suggests directions for future research, including dataset expansion, alternative fine-tuning strategies, and real-time performance evaluation.

Keywords: automatic speech recognition, code-switching, multilingual ASR, low-resource languages

1. Introduction

Automatic Speech Recognition (ASR) has made remarkable progress in recent years, driven by advances in deep learning, the availability of large-scale speech corpora, and increasingly powerful computational resources. Modern ASR systems are now capable of achieving near-human performance on several high-resource languages such as English, Mandarin, and Spanish. These systems have become foundational in applications like voice assistants, transcription services, and accessibility tools.

Despite these achievements, significant challenges remain, particularly in ensuring that ASR technologies are inclusive and perform well across diverse linguistic contexts. Many languages and speech varieties, especially those spoken in Africa and other underrepresented regions, continue to be poorly served due to limited labeled data and resource constraints. This disparity limits the global reach and utility of ASR technologies.

Recent efforts have turned toward multilingual ASR models, such as Whisper (Radford et al., 2022), Canary (Rekesh et al., 2023), OWSM (Peng et al., 2023) and MMS (Pratap et al., 2023), which aim to generalize across multiple languages without requiring language-specific training. While these models show promise, their performance in low-resource and linguistically complex scenarios, such as code-switching (CS), remains underexplored and inconsistent.

Code-switching (CS); the practice of alternating between two or more languages within a single conversation or utterance is a widespread linguistic phenomenon in multilingual communities (Babatunde et al., 2025). In Nigeria, for example, speakers frequently switch between English and indigenous languages such as Yoruba, Igbo, Hausa in natural speech. This practice is particularly prevalent in urban settings, where it serves purposes such as enhancing clarity, expressing social identity, and facilitating ease of communication. Beyond informal discourse, code-switching also plays a critical role in professional domains like healthcare, commerce, and education, where it helps bridge communication gaps and improve engagement. The growing influence of digital communication platforms including social media and messaging apps has further amplified the use of code-switching in both text and voice-based interactions (Abosede and Ayomide, 2021; Babatunde et al., 2025).

Linguistically, code-switching can be categorized into two types: inter-sentential, where language switching occurs between sentence boundaries, and intra-sentential, where the switch happens within a single sentence (Poplack, 2000). These linguistic patterns introduce significant complexity for ASR systems, which must accommodate shifts in phonetics, syntax, and lexical structure often within a single utterance. This complexity makes CS a uniquely challenging task for ASR, especially in low-resource settings where labeled training data is scarce.

2. Related Works

Researchers have explored various methods to improve multilingual ASR models for code-switched (CS) speech. One effective approach involves fine-tuning self-supervised speech representations, such as wav2vec 2.0 XLSR, directly on code-switched data. For instance, Ogunremi et al. (2023) use the South African corpus of multilingual code-switched soap opera speech (Niesler et al., 2018) in their experiments. When combined with augmentation using n-gram language models trained from transcripts, their method reduces absolute word error rates by up to 20% compared to hybrid models trained from scratch on CS data, demonstrating its viability, especially in low-resource settings. Nevertheless, most existing work has primarily focused on multilingual ASR systems or standalone models for languages like Yoruba and English, with relatively little attention given to models that explicitly handle mixed-language input scenarios.

In a related study, fine-tuning both monolingual and multilingual ASR models for Yoruba-English CS speech revealed that unadapted monolingual models can outperform large multilingual models like Whisper Large v3 in zero-shot conditions. Fine-tuning significantly reduced word error rates for both model types, with monolingual models achieving over 20% WER reduction while maintaining lower computational costs. Despite some degradation in English recognition for multilingual models post fine-tuning, monolingual models

offer a computationally efficient and competitive alternative for CS-ASR, particularly in resource-constrained environments (Babatunde et al., 2025).

Although CS ASR has been explored in several high-resource language pairs such as Chinese-English (Lovenia et al., 2021), Mandarin-English (Lyu et al., 2010), and Arabic-English (Mubarak et al., 2021; Ali and Aldarmaki, 2024), African language contexts particularly Yoruba-English code-switching remain significantly underexplored. There is a notable lack of benchmark datasets and comparative evaluations of state-of-the-art ASR models on African CS speech.

This paper addresses this gap by benchmarking and finetuning two multilingual ASR models Whisper and MMS that supports Yoruba and English language independently. We evaluate each model's zero-shot capabilities and their performance after fine-tuning on a domain-specific Yoruba-English code-switching dataset.

3. Methodology

3.1. Dataset

We used a 21-hour Yoruba-English code-switched dataset collected from 24 unique speakers, covering 10 diverse domains: *family, sports, lifestyle, healthcare, business, news, education, agriculture, general,* and *entertainment* (Babatunde et al., 2025). Each utterance is approximately 8 seconds long and contains a mix of Yoruba and English, either in inter-sentential or intra-sentential code-switching fashion.

Table 1: Dataset train/validation/test split

Split	Hours	Percentage	Samples (%)
Train	17.00	80.5	13,121
Validation	2.19	10.4	1,645
Test	1.93	9.1	1,613
Total	21.12	100	16,379

The data set comprises a combination of clean speech and recordings from noisy environments, designed to help the model generalize more effectively and avoid overfitting to a particular speaking style or background condition.

3.1.1. Code-Mixing Analysis

Code-Mixing Index (CMI) is a measure used to quantify the extent or level of code-mixing or code-switching in a given sentence (Chowdhury et al., 2020; Gambäck and Das, 2016; Babatunde et al., 2025). A higher CMI value indicates a greater degree of code-mixing, while a lower value suggests that the sentence or corpus is more monolingual.

Table 2 shows that there are more English-dominant sentences (33.94) in the dataset compared to Yoruba-dominant sentences (32.19). Overall, the dataset has an average CMI of 33.23 indicating that the dataset exhibits a moderate degree of code-mixing.

Table 2: Code-Mixing Index (CMI) statistics by sentence type.

Sentence Type	Avg. CMI	Sentences	Percentage of Total (%)
English-Dominant	33.94	9,327	57.0
Yoruba-Dominant	32.19	7,052	43.0
Overall	33.23	16,379	100

3.2. ASR Models

For this experiment, two multilingual ASR models were selected `Whisper Large v3` and `MMS 1B All`. These models were chosen because they support Yoruba & English independently and their robust multilingual capabilities.

3.2.1. Whisper large V3

Whisper Large v3 is a 1.55 billion-parameter transformer-based encoder-decoder model built on a sequence-to-sequence architecture, allowing it to generate transcriptions token by token (Radford et al., 2022). It was trained on 680,000 hours of multilingual and multitask-supervised data.

Compared to earlier versions, the v3 model processes spectrogram inputs with 128 Mel frequency bins (up from 80) and introduces improved language support, including a dedicated language token for Cantonese.

The model natively supports multiple languages and can perform English-Yoruba transcription out-of-the-box. However, it shows limitations when handling code-switched utterances, as it relies heavily on the initial tokens to determine the language context.

For our experiment, we initialized from the publicly available Whisper Large v3 checkpoint on Hugging Face and fine-tuned it on our code-switched dataset [1].

3.2.2. MMS-1B-ALL

The Massive Multilingual Speech (MMS) model is a 1-billion-parameter ASR model based on the `wav2vec 2.0` architecture (Pratap et al., 2023). It was pre-trainedd on 500,000 hours of unlabeled speech data across 1,406 languages and designed to support transcription in 1,107 languages.

Unlike traditional sequence-to-sequence models, MMS utilizes a conformer-based encoder paired with a Connectionist Temporal Classification (CTC) output layer (Graves, 2012). This design enables more efficient training and inference. Additionally, the model uses language-specific adapters to transcribe over 1,000 languages, which has been shown to be both memory efficient and effective, particularly for low-resource languages.

We used the `facebook/mms-1b-all` checkpoint, which was fine-tuned from MMS-1B on 1,162 languages, including Yoruba [2].

1. https://huggingface.co/openai/whisper-large-v3
2. https://huggingface.co/facebook/mms-1b-all

3.3. Fine-tuning Strategy

In this study, the fine-tuning process was conducted to adapt the multilingual ASR models (`Whisper Large v3` and `MMS-1B-All`) for recognizing Yoruba-English code-switched speech. Below, we describe the fine-tuning strategy employed in detail.

3.3.1. Dataset Preprocessing

Before training, the text data was preprocessed by converting all characters to lowercase and removing punctuation marks, except for intra-word apostrophes and Yoruba diacritics. Language identification tags were not provided to the model at any time.

Resampling: The original audio dataset was resampled to a consistent 16 kHz sampling rate to ensure compatibility with the input requirements of both models. This uniform sampling frequency is essential for effective feature extraction and stable training performance.

Dataset Conversion: The dataset was converted into the Hugging Face `datasets` format, allowing for seamless integration into the training pipeline. This format supports efficient handling of paired audio and text data, and simplifies the processes of loading, batching, and preprocessing.

Tokenizers and Adapters: Both models were trained using tokenization strategies suited to their architectures. `Whisper Large v3` employed a subword-based tokenizer, which effectively handled the complexities of Yoruba-English code switching. In contrast, `MMS-1B-All` utilizes an adapter that defines language-specific tokens. To accommodate Yoruba-English code-switching, a new adapter (Eng-Yor) was created by merging the tokens from both the English and Yoruba adapters. These tokenization approaches ensured accurate representation of both languages and smooth handling of code-switch transitions during training.

For fine-tuning, we first performed zero-shot transcription with both `Whisper Large v3` and `MMS-1B-All` on the dataset. Following this, both models were fine-tuned on the dataset to evaluate improvements in performance. The fine-tuning process helped enhance transcription accuracy, particularly for the more complex code-switching patterns between Yoruba and English. Table 3 contains information about the resource compute that was used in the process of finetuning.

Table 3: Compute Resources for Model Training

Resource	**Details**
GPU Model	NVIDIA Tesla P100
Number of GPUs	1
GPU RAM	16 GB HBM2
Framework Used	PyTorch

4. Results and Discussion

The goal of this study was to benchmark the performance of multilingual ASR models; `Whisper Large v3` and `MMS-1B-All` on an English-Yoruba code-switched dataset, with a focus on improving transcription accuracy. Both models were fine-tuned and evaluated using Word Error Rate (WER) as the primary metric.

Table 4: WER for unfinetuned (UFT) and finetuned (FT) ASR models on code-switched, English-Yoruba test sets

Models	**WER**
UFT MMS-1B-ALL with Eng adapter	0.7257
UFT MMS-1B-ALL with Yor adapter	0.8090
UFT Whisper-v3	**0.6684**
FT MMS-1B-ALL with Eng-Yor adapter	0.3218
FT Whisper-v3	**0.3335**

The following improvements were observed after fine-tuning:

- **MMS-1B-All** achieved a 55.8% reduction in Word Error Rate (WER), improving from 0.7257 to 0.3218.
- **Whisper Large v3** achieved a 50.1% reduction in WER, improving from 0.6684 to 0.3335.

4.1. Comparison of Performance

Table 5: Translation samples before and after fine-tuning for Whisper and MMS models

Samples	sample 1	sample 2
References	o ti clear pe printer yen ko function properly o need repair	drama series yen end in suspense ko clear
Whisper (Unfinetuned)	If you clear up a printer that doesn't function properly you need repair	Drama series yen end in suspense ko clear
MMS-1B-ALL (Unfinetuned Eng Adapter)	o ti cla pe printer yen ko fonction properly o nid repier	drama syrisien end in suspense co clar
MMS-1B-ALL (Unfinetuned Yor Adapter)	ó ti clẹ pé printe yẹn kò funksion properly ó nid ripier	drama sirisien end in suspence ko clair
Whisper (Finetuned)	ò ti clear pé printer yen kò function properly ó need repair	drama series yen end in suspense ko clear
MMS-1B-ALL (Finetuned)	o ti clear pe printer yen ko function properly o ned repair	drama series yen end in suspense ko clear

Before fine-tuning: The `MMS-1B-ALL` model with the English adapter achieved a WER of 0.7257, while using the Yoruba adapter slightly worsened performance to 0.8090. Meanwhile, `Whisper Large v3` demonstrated a lower baseline WER of 0.6684. Sample translations before fine-tuning are provided in Table 5, showcasing how both models performed with raw code-switched data.

After fine-tuning: `MMS-1B-ALL` achieved the lowest WER of 0.3218, outperforming `Whisper Large v3`, which attained a WER of 0.3335. These results indicate that `MMS-1B-ALL`, with its CTC-based architecture, handled the code-switching task slightly better than `Whisper Large v3`. Post-fine-tuning translation samples, shown in Table 5, further highlight the improvements in transcription accuracy.

4.2. Factors Contributing to Performance

Model Architecture: The `MMS-1B-ALL`, as a CTC-based encoder-only model, may have had an advantage in handling code-switched speech. CTC models often better manage shorter sequences and mixed languages since they do not depend on generating entire sequences autoregressively, unlike `Whisper Large v3`'s decoder.

Fine-Tuning Effectiveness: Fine-tuning enhanced both models' ability to capture linguistic transitions between English and Yoruba which is critical for code-switched speech recognition. `MMS-1B-ALL` demonstrated a more substantial WER improvement, likely benefiting from its pre-trained multilingual architecture and the adapter tuning approach.

Dataset and Language Mixing: Although the 21-hour dataset was relatively small, its diverse mix of speech styles and topics enabled effective model adaptation. Both models achieved over 50% reduction in WER, highlighting the positive impact of fine-tuning despite the limited dataset size.

Both models showed significant improvements after fine-tuning, demonstrating the effectiveness of adapting to a domain-specific code-switched dataset. The results indicate that fine-tuning with a relatively small dataset can lead to a substantial reduction in transcription error rates.

5. Conclusion

The results of this study demonstrate that fine-tuning `Whisper Large v3` and `MMS-1B-All` on a Yoruba-English code-switched dataset leads to substantial improvements in transcription accuracy. Specifically, `MMS-1B-All` achieved a **55.8%** reduction in Word Error Rate (WER), while `Whisper Large v3` achieved a **50.1%** reduction.

Although `MMS-1B-All` slightly outperformed `Whisper-v3`, the performance gap was marginal, indicating that both models are viable options for deployment in real-world Yoruba-English code-switched speech recognition applications.

These results confirm that even with a relatively limited dataset of just 21 hours, fine-tuning can yield significant gains in multilingual ASR performance, particularly in low-resource and code-switched language contexts.

6. Future Works

While this study demonstrates promising results, several avenues remain to be explored. First, expanding the dataset to include a larger number of speakers and a wider variety of topics would likely improve model robustness and generalization.

Additionally, evaluating these models on other African language pairs or on global code-switched speech datasets could provide insights into their adaptability across different multilingual contexts.

Exploring alternative fine-tuning approaches, such as combining multiple adapters or employing multi-task learning strategies, may further enhance transcription accuracy.

Finally, assessing model performance in real-time scenarios and evaluating computational efficiency, especially in resource-constrained environments is crucial to facilitate practical deployment and usability.

References

Otemuyiwa Abosede and Iyanuoluwa Ayomide. Effects of code-switching on the acquisition of the english language by english and yoruba language bilinguals. *OLATEJU IA*, page 9, 2021.

Maryam Al Ali and Hanan Aldarmaki. Mixat: A data set of bilingual emirati-english speech. *arXiv preprint arXiv:2405.02578*, 2024.

Oreoluwa Boluwatife Babatunde, Victor Tolulope Olufemi, Emmanuel Bolarinwa, Kausar Yetunde Moshood, and Chris Chinenye Emezue. Beyond monolingual limits: Fine-tuning monolingual asr for yoruba-english code-switching. In *Proceedings of the 7th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 18–25, 2025.

Shammur A Chowdhury, Younes Samih, Mohamed Eldesouki, and Ahmed Ali. Effects of dialectal code-switching on speech modules: A study using egyptian arabic broadcast speech. In *Interspeech*, pages 2382–2386, 2020.

Björn Gambäck and Amitava Das. Comparing the level of code-switching in corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1850–1855, 2016.

Alex Graves. Connectionist temporal classification. In *Supervised sequence labelling with recurrent neural networks*, pages 61–93. Springer, 2012.

Holy Lovenia, Samuel Cahyawijaya, Genta Indra Winata, Peng Xu, Xu Yan, Zihan Liu, Rita Frieske, Tiezheng Yu, Wenliang Dai, Elham J Barezi, et al. Ascend: A spontaneous chinese-english dataset for code-switching in multi-turn conversation. *arXiv preprint arXiv:2112.06223*, 2021.

Dau-Cheng Lyu, Tien Ping Tan, Engsiong Chng, and Haizhou Li. Seame: a mandarin-english code-switching speech corpus in south-east asia. In *Interspeech*, volume 10, pages 1986–1989, 2010.

Hamdy Mubarak, Amir Hussein, Shammur Absar Chowdhury, and Ahmed Ali. Qasr: Qcri aljazeera speech resource–a large scale annotated arabic speech corpus. *arXiv preprint arXiv:2106.13000*, 2021.

Thomas Niesler et al. A first south african corpus of multilingual code-switched soap opera speech. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Tolulope Ogunremi, Christopher D Manning, and Dan Jurafsky. Multilingual self-supervised speech representations improve the speech recognition of low-resource african languages with codeswitching. *arXiv preprint arXiv:2311.15077*, 2023.

Yifan Peng, Jinchuan Tian, Brian Yan, Dan Berrebbi, Xuankai Chang, Xinjian Li, Jiatong Shi, Siddhant Arora, William Chen, Roshan Sharma, Wangyou Zhang, Yui Sudo, Muhammad Shakeel, Jee weon Jung, Soumi Maiti, and Shinji Watanabe. Reproducing whisper-style training using an open-source toolkit and publicly available data. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 12 2023.

Shana Poplack. Toward a typology of code-switching. *L. WEI (éd.), The bilingualism reader. London, New York: Routeledge*, pages 221–255, 2000.

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. Scaling speech technology to 1,000+ languages. *arXiv*, 2023.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL https://arxiv.org/abs/2212.04356.

Dima Rekesh, Nithin Rao Koluguri, Samuel Kriman, Somshubra Majumdar, Vahid Noroozi, He Huang, Oleksii Hrinchuk, Krishna Puvvada, Ankur Kumar, Jagadeesh Balam, et al. Fast conformer with linearly scalable attention for efficient speech recognition. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE, 2023.