Beyond Monolingual Limits: Fine-Tuning Monolingual ASR for Yoruba-English Code-Switching

Oreoluwa Babatunde[1], Victor Olufemi[1], Emmanuel Bolarinwa[1], Kausar Moshood[1], Chris Emezue[2,3]
[1]LyngualLabs, [2]Lanfrica, [3]Mila-Quebec Institute
{oreoluwa, victor, emmanuel, kausar}@lynguallabs.org
chris.emezue@lanfrica.com

Abstract

Code-switching (CS) presents a significant challenge for Automatic Speech Recognition (ASR) systems, particularly in low-resource settings. While multilingual ASR models like OpenAI Whisper Large v3 are designed to handle multiple languages, their high computational demands make them less practical for real-world deployment in resource-constrained environments. In this study, we investigate the effectiveness of fine-tuning both monolingual and multilingual ASR models for Yoruba-English CS speech. Our results show that unadapted monolingual ASR models outperform Whisper Large v3 in a zero-shot setting on CS speech. Fine-tuning significantly reduces WER for both monolingual and multilingual models, with monolingual models achieving over a 20% WER reduction on CS and Yoruba speech while maintaining lower computational costs. However, we observe a trade-off, as fine-tuning leads to some degradation in English recognition, particularly for multilingual models. Our findings highlight that while multilingual models benefit from fine-tuning, monolingual models provide a computationally efficient and competitive alternative for CS-ASR, making them a viable choice for resource-constrained environments.

1 Introduction and Related Works

Automatic Speech Recognition (ASR) has become a vital component of Natural Language Processing (NLP) and speech technology, enabling machines to understand and transcribe spoken language. While early ASR systems were designed for single languages, real-world linguistic environments are far more complex, as people naturally switch between multiple languages. To address this, multilingual ASR systems have been developed, allowing speech recognition across multiple languages within a single model. State-of-the-art models such as OpenAI's Whisper (Radford et al., 2022)and Meta's MMS (Pratap et al., 2023) have demonstrated impressive multilingual capabilities. However, these systems face significant challenges, particularly in handling code-switching (CS)—a phenomenon where speakers alternate between languages within a conversation or an utterance. This is a crucial issue in speech technology due to its prevalence in multilingual communities.

CS is common in bilingual and multilingual communities, such as Nigeria, where over 200 languages are spoken alongside English as the lingua franca. Studies show that over 60% of Nigerians frequently switch between their native languages and English, especially in informal conversations, workplaces, and social media interactions (Abosede and Ayomide, 2021). This phenomenon is driven by Nigeria's multilingual landscape, where English serves as the official language while indigenous languages such as Yoruba, Igbo, and Hausa remain central to daily communication. Code-switching is particularly prevalent in urban areas, where speakers alternate between languages for clarity, social identity, and ease of expression. Beyond social interactions, code-switching is also widely used in healthcare, business, and economic transactions, allowing speakers to bridge communication gaps, ensure better understanding, and facilitate more effective engagement in professional and commercial settings. Additionally, digital platforms, including social media and messaging apps, have further amplified the use of code-switched speech in both text and voice communication.

Code-switching (CS) can be categorized into inter-sentential CS, where language switching occurs at sentence boundaries, and intra-sentential CS, where switching happens within a single sentence (Poplack, 1980). Researchers have explored various methods to improve multilingual ASR models for CS speech; however, these approaches often introduce additional model complexity and higher computational costs. One such approach involves

Proceedings of the 7th Workshop on Computational Approaches to Linguistic Code-Switching, pages 18–25
May 3, 2025 ©2025 Association for Computational Linguistics

fine-tuning the MMS model, a multilingual ASR model with adapters for over 1,000 languages, using techniques like Transformer Code Switching (TCS) and Post Adapter Code Switching (PACS). These techniques integrate language adapters for both matrix and embedded languages to improve CS transcription accuracy (Kulkarni et al., 2023). While effective, they require substantial computational resources due to the large size of multilingual ASR models. Openai Whisper Multilingual Small (244M parameters) and MMS Medium (300M parameters) (Radford et al., 2022; Pratap et al., 2023) are significantly larger than monolingual models, many of which have fewer than 100M parameters. Even the smallest versions of advanced multilingual ASR models, such as Seamless M4T Medium (1.2B parameters) (Barrault et al., 2023), remain computationally large, making fine-tuning for CS tasks more challenging. The sheer size of these models results in higher computational requirements, longer training times, and greater memory usage. Moreover, multilingual ASR models must rapidly adapt between multiple languages within an utterance, requiring an intricate balance of language-specific features, which becomes even more difficult with larger models. These challenges are further exacerbated in low-resource settings, where access to high-performance computing infrastructure is limited, making it difficult to fine-tune and deploy such models effectively.

Another major challenge in enhancing CS-ASR is the scarcity of labeled CS speech data. In low-resource settings, particularly for language pairs like Yoruba-English, Igbo-English, and Hausa-English, the lack of sufficient labeled data significantly impedes ASR models' ability to generalize effectively on code-switched speech. Without adequate training data, these models struggle to learn the diverse patterns of code-switching that naturally occur between languages in speech, leading to poor performance. Ogunremi et al. (2023a) address this challenge by fine-tuning self-supervised models, such as wav2vec 2.0 XLSR, on South African CS speech data, achieving a 20% reduction in word error rates (WER) compared to baseline models trained from scratch. This approach demonstrates that self-supervised pre-training can enhance model performance even when labeled data is scarce. However, it remains resource-intensive, requiring significant computational resources for fine-tuning and careful hyperparameter tuning. A study on Frisian-Dutch CS-ASR explored the use of multilingual deep neural networks (DNNs) with a two-step training process: (1) pretraining on multilingual speech data, including both the target language and related high-resource languages, and (2) retraining the shared hidden layers on a smaller Frisian-Dutch dataset to better adapt the model to code-switched speech. While this approach improved ASR performance, it introduced challenges, such as a reliance on high-resource languages and increased computational demands due to the multi-stage retraining process (Yılmaz et al., 2016).

Several studies have explored CS-ASR by adapting multilingual or monolingual models. In one approach, the bi-encoder structure (Song et al., 2022), fuses two monolingual ASR models for language-specific predictions, combining outputs in a two-stage process: Speech Awareness (SA) and Language Fusion (LF). This method improves efficiency by reducing reliance on large CS datasets and was effective on a Mandarin-English CS corpus.

Model	WER	Params (M)
OpenAI Whisper Large v3	0.6684	1550
FastConformer CTC Large	0.6473	120
Conformer CTC Large	0.6469	118.8
FastConformer Transducer Large	**0.6294**	120

Table 1: Zero-shot WER comparison of unadapted monolingual ASR models and OpenAI Whisper Large v3 on Yoruba-English CS speech.

As shown in Table 1, OpenAI Whisper Large v3, despite being designed for multiple languages, including both Yoruba and English, exhibits the highest WER on Yoruba-English CS speech when evaluated in a zero-shot setting using the CS test set. Importantly, none of the models in this evaluation, including Whisper and the monolingual ASR models, have been fine-tuned on CS data. Beyond its higher WER, Whisper's large parameter size (1.55 billion) results in significantly greater computational demands. In contrast, the monolingual ASR models, with approximately 120 million parameters, achieve lower WERs while offering substantial advantages in efficiency and resource requirements.

These findings suggest that fine-tuned monolingual models offer a computationally efficient and high-performing alternative for CS-ASR in low-resource environments. While multilingual ASR models like Whisper large v3 provide broad language coverage, their high computational demands and inference latency make them less practical for real-world deployment in resource-constrained settings.

Given these considerations, we propose fine-tuning a monolingual ASR model—originally trained on English—to efficiently recognize Yoruba-English CS speech. This approach balances performance and computational efficiency, enabling ASR systems that are both accurate and deployable on low-resource hardware.

To assess the viability of this approach,we investigate the following key research questions:

1. **Adaptability to Code-Switching:** Can a fine-tuned monolingual ASR model effectively recognize and transcribe speech that mixes English and Yoruba while maintaining a computational advantage over multilingual models?

2. **Recognition of Yoruba-Specific Speech:** Given that the base model was trained on English, how well can it learn Yoruba-specific phonetics, vocabulary, and linguistic structures while remaining computationally efficient?

3. **Retention of English Proficiency and Catastrophic Forgetting:** Does fine-tuning for code-switching degrade the model's performance on English-only speech, or can it retain its original proficiency while improving CS transcription?

4. **Performance vs. Computational Trade-offs:** How does the trade-off between WER improvements and computational demands differ between fine-tuned monolingual models and multilingual models like Whisper Large v3? What are the implications for ASR deployment in low-resource settings?

2 Monolingual ASR for Yoruba-English Code-Switching

Monolingual models for code-switched ASR are relatively underexplored, as most research has focused on multilingual or hybrid models (e.g., bi-encoders) that handle multiple languages simultaneously (Radford et al., 2022; Pratap et al., 2023; Mustafa et al., 2022; Kulkarni et al., 2023; Barrault et al., 2023; Ogunremi et al., 2023a; Yılmaz et al., 2016; Song et al., 2022). Monolingual models offer a computationally efficient alternative, particularly in resource-constrained settings.

Much of the CS-ASR research has concentrated on high-resource language pairs such as Chinese-English (Lovenia et al., 2021), Mandarin-English (Lyu et al., 2010), and Arabic-English (Ali and Aldarmaki, 2024; Mubarak et al., 2021), leveraging large datasets and advanced models. In contrast, research on African language CS-ASR specifically Yoruba-English remains untouched.

The Yoruba language is spoken in several West African countries, including Nigeria, Benin Republic, and parts of Togo and Sierra Leone, making it one of the largest single languages in sub-Saharan Africa. Additionally, Yoruba is spoken in diaspora communities, particularly in Cuba and Brazil. Beyond these regions, Yoruba people are among the most traveled African ethnic groups, often settling in the United States, the United Kingdom, and other parts of Europe. In these environments, they tend to live in close-knit communities, where code-switching between Yoruba and English becomes a sine qua non in daily interactions. This widespread usage underscores the significance of studying Yoruba-English code-switching for ASR development.

Furthermore, only a few code-switched speech datasets exist for African languages, with most research focusing on South African language pairs such as English-Zulu (Eng-Zul), English-Xhosa (Eng-Xho), English-Sotho (Eng-Sot), and English-Tswana (Eng-Tsn) (Ogunremi et al., 2023b). The lack of resources and dedicated research on Yoruba-English CS-ASR presents a significant gap in the field.

3 Experimental Setup

This section presents the dataset, the selected models, and the fine-tuning strategy used in our experiments.

3.1 Data

The data used in this study consists of 21 hours of transcribed Yoruba-English code-switched speech from 24 unique speakers. The dataset ensures diversity in accents and speaking styles while capturing both inter-sentential (switching between sentences) and intra-sentential (switching within a sentence) code-switching patterns. To enhance model robustness, it includes a balanced mix of clean and noisy recording conditions. The average utterance length is 8 seconds.

To ensure broad linguistic and contextual representation, the dataset spans 10 diverse domains, including *family, sports, lifestyle, healthcare, business, news, education, agriculture, general, and entertainment*.

This dataset is part of an ongoing collection effort aimed at reaching 100 hours of annotated Yoruba-English code-switched speech data. To reproduce this research the 21 hours data can be found here[1]. However, once the target of 100 hours is reached, the full dataset will be released on Hugging Face to support research in code-switching ASR and ensure long-term accessibility for the research community.

Split	Hours	Percentage (%)	Samples
Training	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

Table 2: Dataset split for training, validation, and testing.

Table 2 presents the dataset split used for fine-tuning. The training set comprises 80.5% of the total 21.12-hour dataset, while the validation and test sets account for 10.4% and 9.1%, respectively. This split ensures ample training data while preserving robust evaluation metrics. The test set utterances were entirely excluded from the training and validation sets. Although there was speaker overlap between the training and validation sets, the test set comprised only entirely unseen speakers, providing a reliable measure of generalization.

3.2 Code-Switching Analysis

To quantify the extent of code-mixing in a given sentence, we use the Code-Mixing Index (CMI) (Chowdhury et al., 2020), which is defined as:

$$CMI^i = w_N \left(\frac{\min(N_y^i, N_e^i)}{N^i} \right) + w_\alpha \frac{\alpha^i}{N^i} \quad (1)$$

where:

- N^i is the total number of words in the i-th sentence,
- N_y^i and N_e^i represent the number of words in Language y(Yoruba) and Language e (English), respectively, in the i-th sentence,
- α^i is the number of code-switching points in the i-th sentence,
- w_N and w_α are weight parameters (both set to 0.5 in our implementation).

The term $\frac{\min(N_y^i, N_e^i)}{N^i}$ captures the degree of balance between the two languages in the sentence, ensuring that higher values indicate more intermixing. The second term, $\frac{\alpha^i}{N^i}$, accounts for the frequency of code-switching points. The weights w_N and w_α control the relative contribution of these two factors.

A higher CMI value indicates a greater degree of code-mixing, while a lower value suggests that the sentence is more monolingual.

3.2.1 Sentence Classification Based on Dominant Language

To better understand the nature of code-switching in our dataset, we categorize sentences based on their dominant language, which is determined by the majority language of tokens in each utterance:

- **English-Dominant Sentence:** A sentence in which English constitutes the majority of tokens, with Yoruba words appearing as insertions.
- **Yoruba-Dominant Sentence:** A sentence where Yoruba is the primary language, but it includes insertions from English.

The classification allows us to analyze whether code-switching is more prominent when speakers primarily use Yoruba or English.

Sentence Type	Avg. CMI	Sentences
English-Dominant	33.94	9,327
Yoruba-Dominant	32.19	7,052
Overall	**33.23**	**16,379**

Table 3: Code-Mixing Index (CMI) statistics by sentence type.

The overall average CMI for our dataset is 33.23, indicating a moderate degree of code-mixing across English and Yoruba. The slightly higher CMI for English-dominant sentences (33.94) compared to Yoruba-dominant ones (32.19) suggests that speakers tend to integrate more words from the dominant language when mixing. These findings highlight the linguistic complexity of our dataset, reinforcing the need for ASR models capable of handling mixed-language utterances effectively. The observed code-mixing patterns also provide insights into language dominance shifts, which can inform the development of better multilingual and code-switching ASR systems.

[1] You can access the dataset here: Data.

4 ASR Models

For our experiments, we evaluated a range of ASR models, including both monolingual and multilingual models, as well as their fine-tuned versions, on Yoruba-English code-switched speech. We selected three monolingual ASR models from NVIDIA's STT (Speech-to-Text) series, which are some of the best-performing models on the open ASR leaderboard on Hugging Face.[2]. These models include:

- **fastconformer_ctc_large**: A Conformer-based model optimized with CTC loss for efficient speech recognition (Rekesh et al., 2023).
- **conformer_ctc_large**: A variant designed for enhanced ASR performance, utilizing the Conformer architecture (Gulati et al., 2020) .
- **fastconformer_transducer_large**: A faster version that incorporates Transducer loss, suitable for real-time applications (Rekesh et al., 2023).

We fine-tuned these monolingual models on our Yoruba-English code-switched dataset to adapt them for code-switching speech. This fine-tuning was aimed at enabling the models to recognize both Yoruba and English phonetics, tonal variations, and mixed-language structures. Additionally, we fine-tuned **OpenAI Whisper large v3**, a state-of-the-art multilingual ASR model, on Yoruba-English code-switched speech. Since Whisper was pretrained on a large multilingual corpus that includes English and Yoruba, we sought to determine if this prior exposure could enhance its ability to transcribe code-switched speech compared to the monolingual models. The fine-tuning of both monolingual and multilingual models involved adapting them to handle spontaneous code-switching in Yoruba-English speech, with specific strategies tailored to each model's architecture.

4.1 Fine-tuning Monolingual and Multilingual ASR Models

We fine-tuned both monolingual and multilingual ASR models on our Yoruba-English code-switched dataset. For monolingual models, we adapted pre-trained English-only models, which lacked exposure to Yoruba phonetics and mixed-language structures. Fine-tuning included training a new SentencePiece tokenizer, using their respective loss functions (CTC or transducer loss), and adapting the models to the combined Yoruba-English dataset. For the multilingual Whisper Large v3 model, fine-tuning focused on improving its ability to handle code-switching. We fine-tuned the model using its default sequence-to-sequence loss, optimizing both encoder and decoder components for better mixed-language speech recognition.

Model	Params (M)	Decoder	Type
nvidia/conformer_ctc_large	118.8	CTC	Mono
nvidia/fastconformer_ctc_large	120.0	CTC	Mono
nvidia/fastconformer_transducer_large	120.0	RNN-T	Mono
openai/whisper-large-v3	1550.0	Seq2Seq	Multi

Table 4: ASR Models Used in Our Experiments

Table 4 provides details on the ASR models used. The Nvidia Conformer and FastConformer models with CTC decoders predict sequences frame-independently, while the FastConformer model with a transducer (RNN-T) decoder is designed for streaming ASR. Unlike these models, OpenAI Whisper v3 employs an encoder-decoder Transformer architecture, where the encoder processes input audio into a latent representation, and the decoder autoregressively generates text tokens. The decoder uses cross-attention to incorporate contextual dependencies across entire sequences, enabling accurate transcriptions, particularly in code-switched and multilingual scenarios.

Resource	Specification
GPU Model	NVIDIA RTX 6000 Ada
Number of GPUs	1
Memory (RAM)	48GB
Framework	PyTorch + NeMo

Table 5: Compute Resources Used for Fine-Tuning

Table 5 presents the compute resources used for fine-tuning, including the training hyperparameters and time spent for fine-tuning. The fine-tuning process was conducted on a single NVIDIA RTX 6000 Ada GPU with 48GB of memory. The NeMo framework, built on PyTorch, was utilized for efficient model training.

*Note:*The monolingual models were trained for 50 epochs in approximately 11 hours, while the multilingual Whisper model was trained for only 10 epochs over 12 hours. Training Whisper for 50 epochs would have been impractical due to its significantly larger size and computational demands. Additionally, we observed signs of catastrophic forgetting after extended fine-tuning (as evidenced by

[2] https://huggingface.co/spaces/hf-audio/open_asr_leaderboard

Hyperparameter	Value	Training Time
Batch Size	16	-
Learning Rate	1.0×10^{-6}	-
Optimizer	AdamW	-
Monolingual Models		
Number of Epochs	50	11 hours (50 epochs)
Loss Function	CTC / RNN-T	-
Multilingual Whisper Large v3		
Number of Epochs	10	12 hours (10 epochs)

Table 6: Training Hyperparameters and Time for Fine-Tuning

Table 7), which led us to conclude that 10 epochs was an optimal stopping point to preserve its multilingual capabilities while improving code-switched ASR performance.

ASR Model	CS	Yor	Eng
Unfinetuned Models			
fastconformer_ctc_large	0.6473	1.0531	0.1647
conformer_ctc_large	0.6469	1.0516	0.1660
fastconformer_transducer_large	**0.6294**	1.0347	0.1423
Whisper large v3 (Multilingual)	0.6684	**1.0222**	**0.1299**
Finetuned Models			
fastconformer_ctc_large	0.3340	0.8339	**0.4089**
conformer_ctc_large	0.3414	**0.8157**	0.4592
fastconformer_transducer_large	**0.1481**	0.8212	0.5342
Whisper large v3 (Multilingual)	0.3335	0.9859	0.5860

Table 7: WER for unfinetuned and finetuned ASR models on code-switched, Yoruba, and English test sets.

5 Experimental Results

This section presents the results of fine-tuning monolingual and multilingual ASR models for English-Yoruba code-switching, addressing our research questions. We evaluate performance using three test sets: code-switched (CS), Yoruba-only (Yor), and English-only (Eng) speech. The Yoruba test set is from OpenSLR[3], and the English test set is from OpenSLR[4], both providing high-quality speech data for ASR evaluation. We assess monolingual models, including FastConformer and Conformer, as well as multilingual Whisper Large v3. While monolingual models focus on single-language speech, multilingual models leverage cross-lingual knowledge, making them suitable for code-switching. We use Word Error Rate (WER) to measure transcription accuracy based on word substitutions, deletions, and insertions.

6 Discussion

Our study evaluates the effectiveness of fine-tuning monolingual and multilingual ASR models for Yoruba-English code-switching (CS) while prioritizing computational efficiency. We analyze four key aspects:

6.1 Adaptability to Code-Switching

Table 7 shows that un-finetuned monolingual ASR models struggle with code-switched speech due to their English-only training. However, after fine-tuning, their WER on CS speech drops significantly—demonstrating that exposure to CS data enables monolingual models to effectively transcribe mixed-language utterances.

Whisper Large v3, despite being a multilingual model trained on both English and Yoruba, initially performs worse than some monolingual models in recognizing CS speech, with an un-finetuned WER of 0.6684. This suggests that general multilingual training does not automatically confer strong code-switching capabilities. However, after fine-tuning, Whisper Large v3 achieves a WER of 0.3335, making it competitive with the best-performing monolingual models.

Critically, Whisper Large v3's improved CS transcription comes at a significantly higher computational cost, requiring more processing power during both training and inference. This makes fine-tuned monolingual models a more practical choice for low-resource environments, where computational efficiency is paramount.

6.2 Recognition of Yoruba-Specific Speech

Un-finetuned monolingual ASR models perform poorly on Yoruba speech, with WER values around 1.05, as expected due to their lack of exposure to Yoruba phonetics, tones, and linguistic structures. Fine-tuning significantly improves Yoruba recognition, reducing WER to 0.8212 for `fastconformer_transducer_large`. Whisper Large v3, which has seen Yoruba during pretraining, starts with a slightly better WER (1.0222) but still requires fine-tuning for optimal recognition. However, after fine-tuning, monolingual models outperform Whisper Large v3 on Yoruba speech, suggesting that domain-specific adaptation is more effective than multilingual pretraining for handling Yoruba's unique linguistic features. Despite these gains, WER remains relatively high for Yoruba speech across all models, indicating that additional Yoruba-language data could further improve ASR accuracy.

[3] https://openslr.org/86/

[4] https://openslr.org/70/

6.3 Retention of English Proficiency and Catastrophic Forgetting

Table 7 shows that fine-tuning improves CS and Yoruba recognition but leads to performance degradation on English-only speech. After fine-tuning, the WER on English speech increases from 0.16 to 0.41–0.53 for monolingual models and from 0.1299 to 0.586 for Whisper Large v3 The sharper decline in Whisper Large v3's English accuracy suggests that multilingual models may be more susceptible to catastrophic forgetting, as fine-tuning on CS speech shifts their linguistic distribution away from English. This trade-off must be considered when adapting ASR models for multilingual or CS applications.

6.4 Performance vs. Computational Trade-offs

A major consideration in ASR development is the balance between performance and computational cost. While Whisper Large v3 benefits from large-scale multilingual pretraining, its significantly higher resource requirements make it impractical for many real-world applications.

Model	WER (CS)	Time (s)	GFLOPs/sec
fastconformer_ctc_large	0.3340	0.26	2.78
conformer_ctc_large	0.3414	0.56	8.04
fastconformer_transducer_large	0.1481	1.57	2.63
Whisper Large v3	0.3335	1.98	1295.75

Table 8: WER vs. Inference Time and GFLOPs for Finetuned Models.

Table 8 highlights that while Whisper Large v3 and monolingual models achieve similar WER after fine-tuning, monolingual models are significantly faster and require far fewer computational resources. GFLOPs (Giga Floating Point Operations per Second) measure how many billion calculations a model performs per second. Whisper Large v3's extremely high GFLOPs/sec value suggests a substantial increase in processing demands, making it less feasible for deployment in real-time or resource-constrained environments.

In contrast, FastConformer-based models offer a more efficient trade-off between accuracy and computational cost, making them a practical choice for applications requiring low-latency processing and reduced computational overhead.

6.5 Key Takeaways

Our findings highlight several critical insights for CS-ASR:

- Fine-tuned monolingual models can achieve comparable or superior performance to Whisper Large v3 on CS and Yoruba speech while maintaining significantly lower computational costs.
- Inference efficiency is a major bottleneck for Whisper Large v3, making monolingual models a more practical alternative for real-time ASR in low-resource settings.
- Fine-tuning monolingual models on CS data enables effective adaptation to Yoruba phonetics and mixed-language speech, even though some English degradation occurs.
- Multilingual pretraining does not inherently optimize for CS speech, reinforcing the need for domain-specific fine-tuning.

7 Conclusion

Our results show that while large-scale multilingual models like Whisper v3 are designed for cross-lingual speech recognition, their computational cost makes them impractical for real-time, low-resource CS-ASR systems. Instead, fine-tuning monolingual ASR models provides a computationally efficient alternative that achieves competitive performance on code-switched speech while maintaining lower inference latency and hardware requirements. Future research should explore more efficient multilingual adaptation techniques that balance accuracy and computational efficiency.

7.1 Future Works

Future research should explore hybrid approaches, such as combining the efficiency of monolingual models with selective fine-tuning of multilingual models, to optimize both WER and inference efficiency. This suggests that Whisper's multilingual architecture is more susceptible to shifts in linguistic focus after fine-tuning, leading to greater loss in its original English proficiency compared to monolingual models. This is a key trade-off that must be considered when adapting large-scale multilingual models for specific code-switched domains.

Limitations

Our study highlights the effectiveness of monolingual ASR models for Yoruba-English CS speech, but limitations remain. Fine-tuning leads to catastrophic forgetting, increasing WER on English-only speech. Second, our evaluation is limited to Yoruba-English, and further research is needed to assess the generalizability of these findings to other language pairs. The extent to which monolingual models can adapt to different CS contexts remains an open question. Lastly, data scarcity limits training and evaluation, underscoring the need for larger, more diverse CS datasets.

Acknowledgments

We sincerely thank all contributors and volunteers who have dedicated their time and effort to curating and expanding our growing dataset. Their invaluable work in data collection, annotation, and validation has been instrumental in advancing research on code-switching in automatic speech recognition.

References

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

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

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. 2023. Seamlessm4t-massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

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

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Atharva Kulkarni, Ajinkya Kulkarni, Miguel Couceiro, and Hanan Aldarmaki. 2023. Adapting the adapters for code-switching in multilingual asr. *arXiv preprint arXiv:2310.07423*.

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

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

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

Mumtaz Begum Mustafa, Mansoor Ali Yusoof, Hasan Kahtan Khalaf, Ahmad Abdel Rahman Mahmoud Abushariah, Miss Laiha Mat Kiah, Hua Nong Ting, and Saravanan Muthaiyah. 2022. Code-switching in automatic speech recognition: The issues and future directions. *Applied Sciences*, 12(19):9541.

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

Tolulope Ogunremi, Kola Tubosun, Anuoluwapo Aremu, Iroro Orife, and David Ifeoluwa Adelani. 2023b. \{I} r\{o} y\{i} nspeech: A multi-purpose yor\{u} b\'{a} speech corpus. *arXiv preprint arXiv:2307.16071*.

Shana Poplack. 1980. Sometimes i'll start a sentence in spanish y termino en espanol: toward a typology of code-switching1.

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.

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

Tongtong Song, Qiang Xu, Haoyu Lu, Longbiao Wang, Hao Shi, Yuqin Lin, Yanbing Yang, and Jianwu Dang. 2022. Monolingual recognizers fusion for code-switching speech recognition. *arXiv preprint arXiv:2211.01046*.

Emre Yılmaz, Henk van den Heuvel, and David Van Leeuwen. 2016. Code-switching detection using multilingual dnns. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 610–616. IEEE.