Adapting Foundational ASR Models to Efik: An Empirical Study of an Extremely Low-Resource Tonal Language

Offiong Bassey Edet[1,4], **Stephen Orok Duke**[1], **Enoima Essien Umoh**[1], **Benjamin Okon Nyong**[2], **Andrew Asuquo Nkpanam**[3]
[1]University of Cross River State, [2]Arthur Jarvis University, [3]University of Calabar, [4]ML Collective
offiongbassey99@gmail.com, orokduke2003@unicross.edu.ng,
enoimaumoh@unicross.edu.ng, benokon26@gmail.com, drewsuqi6@gmail.com

Abstract

Automatic Speech Recognition (ASR) has significantly transformed human-computer-interaction and natural language processing. However, many African spoken languages, including Efik, remain severely underrepresented in ASR research. This paper investigates the adoption of state-of-the-art foundational ASR models such as XLS-R and Whisper through fine-tuning for Efik, a low-resource tonal language and empirically evaluates their performance. We curate a 3-hour Efik speech dataset and conduct a comparative evaluation using standard ASR metrics. We further augmented the XLS-R CTC model with a 3-gram KenLM language model trained on an Efik text corpus. Experimental results show that XLS-R-300M + KenLM achieves a word error rate (WER) of 10.86% and a character error rate (CER) of 3.16%, substantially outperforming both the baseline XLS-R (WER: 29.2%, CER: 6.4%) and Whisper across noisy and multi-speaker conditions. These findings suggest that lightweight CTC models augmented with language model integration offer a more robust and practical approach for extremely low-resource tonal languages than larger sequence-to-sequence models.

Keywords: Automatic Speech Recognition, Low-Resource Language, Efik, Tonal Language, Transfer Learning

1. Introduction

Automatic Speech Recognition (ASR) has significantly advanced Natural Language Processing (NLP), particularly with the emergence of state-of-the-art (SOTA) models. Recent deep learning-based ASR systems, including self-supervised encoder-only models such as XLS-R (Babu et al., 2021) and encoder-decoder architectures such as Whisper (Radford et al., 2022) are trained on hundreds of thousands of speech data that have demonstrated strong performance in transcribing spoken language into text. Beyond transcription, these models have enabled the development of robust virtual assistants, real-time speech-to-speech translation systems and language learning tools.

Despite these advances, the benefit of modern ASR systems have been largely concentrated on high-resource languages such as English and Mandarin (Imam et al., 2025). In contrast, many African languages, spoken by over one billion people (Adelani et al., 2022), remain severely underrepresented. Africa is home to over 2,000 languages (Abbott and Martinus, 2019), with Nigeria alone accounting for more than 1,000 distinct languages, among which Efik is included.

Efik, spoken by 1.5 million native speakers and about 3 million second-language speakers (Mensah and Mensah, 2014), has received little attention in ASR. As a tonal language with morphological complexity and limited publicly available speech resources and linguistic tools, Efik presents unique challenges that have contributed to its underrepresentation in both ASR and NLP research more broadly.

This study investigates the adaptation of powerful SOTA ASR models to Efik and empirically evaluates their performance in a low-resource setting.

2. A Brief Discussion on Efik Language

Efik is a Benue-Congo language belonging to the Niger-Congo family. It is primarily spoken in Cross River State and parts of Akwa Ibom State in south-eastern Nigeria, and is also spoken in the South Western part of Cameroon (Offiong and Ansa, 2013). Efik language is amongst the earliest Nigerian languages to be written and studied in Nigeria. Its orthography, lexicon and grammatical structure were largely developed through early missionary linguistic efforts (Offiong and Ansa, 2013).

As a tonal language, Efik employs pitch variations to distinguish lexical and grammatical meaning, which introduces additional challenges for automatic speech recognition systems. Combined with its morphological complexity and limited availability of publicly accessible speech resources, these characteristics contribute to the underrepresentation of Efik in ASR research.

Proceedings of SPEAKABLE @ LREC 2026, pages 1–7

3. Related Work

Recent studies have explored ASR development for African and other low-resource languages. Rufai et al. (2020) developed an end-to-end ASR system for Nigeria Pidgin English, demonstrating improvement using SOTA models such as Nemo Quartznet, Wav2Vec2.0 Base-100H, and Wav2Vec XLS-R-Large-53, achieving a word error rate (WER) of 28.6% from Wav2Vec2.0 XLS-R-Large-53.

Similarly, Chanie et al. (2023) developed ASR systems for the three East African languages - Kinyarwanda, Swahili and Luganda trained on 3,900 hours of code switched speech data, achieving competitive WERs across all languages.

Benchmarking efforts by Nahabwe et al. (2025) empirically evaluates four SOTA ASR models - Whisper, XLS-R, MMS, and wav2bert - across 13 African languages using a transcribed dataset of approximately 400 hours, highlighting the strength and limitation of each model in low-resource setting.

Other notable ASR research efforts in Africa have focused on major Nigerian languages such as Yoruba, Hausa, and Igbo, supported by relatively larger speech datasets (Tolúlópè Ógúnrèmí, 2024; DeRenzi et al., 2025). However, Efik remains largely absent from these studies.

To the best of our knowledge, no prior work has addressed automatic speech recognition for Efik, making this study the first empirical investigation of Efik ASR.

4. Efik Language Dataset Curation

The dataset used for this study consists of self-recorded speech data collected from a single native Efik speaker. It comprises 2,632 clips with a total duration of approximately 3 hours. The recording materials were drawn primarily from educational texts, folktales and storybooks to ensure linguistic diversity and natural sentence structure.

All recordings were conducted in a quiet environment using a wireless microphone to minimize background noise and ensure consistent audio quality.

4.1. Labeling and Validation of Audio Recordings

Due to the unavailability of existing Efik speech technologies, a hybrid method was adopted for dataset labeling and validation. First, complete audio recordings with their corresponding transcripts were collected. A custom python script was developed to segment the recordings into smaller utterances. The segmentation process relied on short, natural pauses in speech to avoid generating overly long clips, with a maximum clip duration capped at 16 seconds. Additionally, a padding of 120 milliseconds was applied at the beginning of each clip to prevent truncation of initial phonemes and to reduce overly aggressive splitting by the segmentation algorithm.

Second, dataset validation was performed entirely through manual inspection, ensuring that each audio clip aligned with its corresponding transcription. Although forced alignment tools were considered, existing SOTA ASR models including Whisper, wav2vec2.0 and XLS-R - performed poorly on Efik speech and were therefore unsuitable for reliable alignment.

4.2. Statistical Analysis of Dataset

The dataset contains 2,632 utterances with a total duration of 3.08 hours. Clips lengths range from 0.48s to 15.94s with an average duration of 4.21s and a median of 3.49s. In Figure 1, the y-axis represents the number of clips and the x-axis represents the duration in seconds.

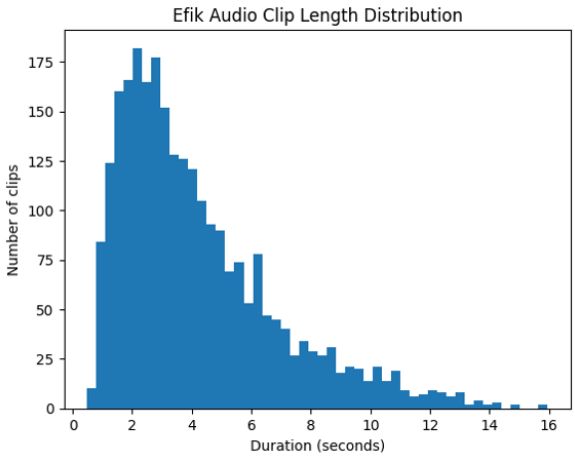

Figure 1: Distribution of audio clips duration.

The corresponding text corpus consists of 23,986 word tokens with a vocabulary size of 3,276 unique words. The average utterance length is 9.11 words, with transcripts ranging from 1 to 33 words. In Figure 2, the y-axis represents the number of clips and the x-axis represents the number of words per utterance.

4.3. Dataset Preprocessing

All transcripts were normalized using Unicode NFC normalization to ensure consistent character representations. Punctuation marks not corresponding to acoustic events were removed. Following standard CTC-based ASR practice, whitespace was replaced with a dedicated word bounding symbols (|) to facilitate alignment and decoding. The dataset

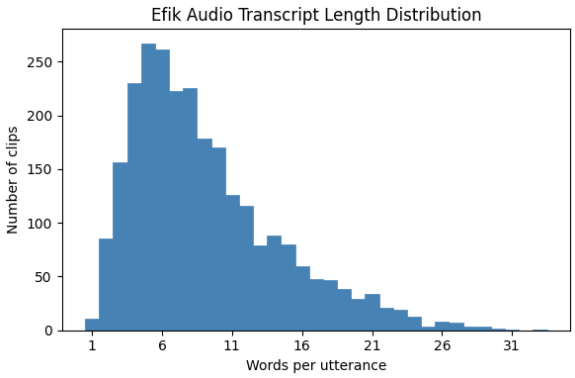

Figure 2: Distribution of transcript length per audio clip.

was partitioned into training and test sets as summarized in Table 1.

Split	Utterances
Train	2,368
Validation	264
Test	393

Table 1: Dataset split statistics.

4.4. Data Availability

We have open-sourced the dataset used in this study. It is publicly available at https://huggingface.co/datasets/offiongbassey/efik_audio_dataset (offiongbassey/efik_audio_dataset).

5. Methodology

5.1. Model Selection

Transfer learning from high resource languages has been widely shown to be an effective approach for improving end-to-end ASR performance in low-resource settings (Xu et al., 2016; Imam et al., 2025; Olatunji et al., 2023). In this study, we select two state-of-the-art (SOTA) ASR models - XLS-R and Whisper for fine-tuning and evaluation of Efik speech.

Both models provide multiple pretrained checkpoints ranging from small to large. Due to limited computational resources and relatively small size of available Efik speech data, we adopt smaller checkpoints for both architectures. Leveraging cross-lingual transfer learning is both theoretically and practically preferable to training from scratch in such low-resource scenarios.

XLS-R (Babu et al., 2021) is a large-scale multilingual speech representation extending wav2vec2.0 (Baevski et al., 2020). It was pretrained on approximately 436,000 hours of unlabeled speech across 128 languages, including several African languages. The pretraining data includes Swahili (91 hours), Yoruba (75 hours), Zulu (56 hours), Lingala (72 hours), Kinyarwand (1,199 hours) and Afrikaans (87 hours), among others, making XLS-R well suited for cross-lingual adaption for African languages.

Whisper (Radford et al., 2022) is an encoder-decoder ASR model trained on 680,000 hours of multilingual and multitask speech data covering approximately 96 languages. Its sequence-to-sequence and large-scale weakly supervised training enable strong robustness to noise, speaker variation and domain mismatch, often requiring minimal fine-tuning for reasonable performance.

5.2. Language Model Integration

CTC-based models such as XLS-R decode speech by independently predicting the most likely token at each time step, without explicit modelling of linguistic context. To address this limitation, we augment the fine-tuned XLS-R model with a 3-gram KenLM language model (Heafield, 2011) trained on an Efik text corpus of 166,977 tokens and 15,662 unique word types. The language model was integrated into the decoding pipeline using pyctcdecode with beam search (beam width=100), a language model weight of $\alpha = 0.5$ and a word insertion bonus of $\beta = 1.0$. This approach allows the decoder to favour linguistically plausible word sequences during inference, without requiring additional annotated speech data.

5.3. Evaluation Metrics

To empirically evaluate the performance of the fine-tuned SOTA ASR models, we used Word Error Rate (WER), a widely adopted metrics for measuring accuracy in ASR systems. WER means the number of substitutions, deletions and insertions required to transform a hypothesis transcription into a reference transcription (Park et al., 2008).

$$\text{WER} = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C} \quad (1)$$

Where S is the number of substitutions; D is the number of deletions; I is the number of insertions; C is the number of correct words; and N is the number of words in the reference (N = S + D + C).

In addition to WER, we report Character Error Rate (CER) which measures the percentage of incorrect characters between the hypothesis and reference text. Unlike WER, which struggles with morphological complex languages, CER has been shown to correlate well with human judgements in

multilingual evaluations and even handles unclear word boundaries (K et al., 2024).

$$\text{CER} = \frac{S + D + I}{N} \quad (2)$$

Where S is the number of substitutions; D is the number of deletions; I is the number of insertions and N is the number of words in the reference.

To assess semantic robustness beyond WER and CER, we adopt a Machine Translation (MT) based evaluation pipeline (Figure 3), where Efik ASR outputs are translated into English and evaluated using BLEU, ChrF and human judgement. This approach better reflects real-world usage, particularly under domain and speaker mismatch.

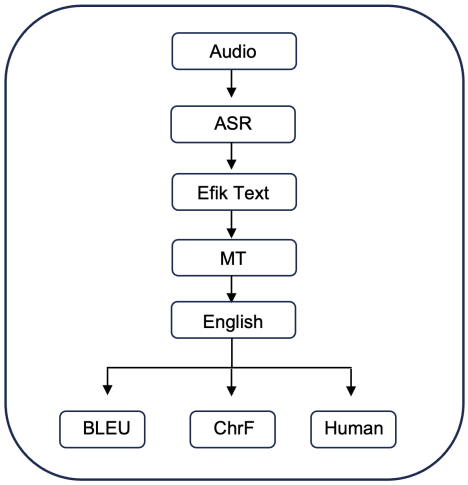

Figure 3: MT - Based Semantic Evaluation Pipeline for Efik ASR

5.4. Experiments

We fine-tuned XLS-R and Whisper using 3 hours of self-recorded Efik speech data, leveraging cross-lingual transfer learning. Each model was trained for 12 epochs.

For XLS-R, we used a learning rate of 3×10^{-4} with a batch size of 4, while Whisper was fine-tuned using a learning rate of 1×10^{-5} with a batch size of 8. All experiments were conducted on NVIDIA A100 with high-ram using the Google Colab Pro+ platform.

The AdamW optimizer was employed with a linear learning rate warm-up of 500 steps. Furthermore, we used the default dropout configuration of the pretrained models during fine-tuning. Mixed precision (FP16) training and gradient checkpoint were enabled to reduce memory usage.

During training, both models exhibited a steady decrease in training and validation loss, indicating progressive learning and effective generalization despite the limited size of the dataset.

6. Results

The results of the study using the two SOTA models - XLS-R and Whisper are presented below:

6.1. Quantitative Analysis

Table 2 presents the performance of both models evaluated on speech from the same speaker in a quiet environment. Both models demonstrate measurable improvements in automatic speech recognition (ASR) for Efik, a low-resource tonal language, despite the limited amount of available training data. The fine-tuned XLS-R-300M model achieves lower WER and substantially lower CER compared to the Whisper baseline. Adding a 3-gram KenLM language model further improves performance, yielding the lowest WER and CER overall.

Model	WER	CER
Whisper	37.5%	28.9%
XLS-R-300M	29.2%	6.4%
XLS-R + KenLM 3-gram	**10.86%**	**3.16%**

Table 2: ASR performance comparison on the Efik test set.

6.2. Qualitative and Semantic Evaluation beyond WER and CER

Quantitative metrics such as WER and CER provide a limited view of transcription quality, particularly concerning semantic preservation. To complement these measures, we conduct a qualitative evaluation across diverse acoustic and speaker scenarios, including single-speaker clean speech and multi-speaker noisy environments. This analysis reveals the interplay between phonetic precision and contextual reasoning, elucidating the trade-offs observed in our automated semantic metrics (BLEU/ChrF).

Table 3 presents representative examples comparing reference transcripts with outputs from Whisper, XLS-R, and XLS-R augmented with a 3-gram KenLM language model, highlighting differences in semantic fidelity and robustness under domain shift. The KenLM-enhanced model demonstrates improved transcription quality in both clean and noisy conditions, producing outputs that more closely align with the reference transcripts, particularly in preserving morphological boundaries and reducing fragmentation.

We further adopt a translation-based semantic evaluation, in which Efik ASR outputs are translated into English and assessed using BLEU and

Scenario	Reference (Source)	XLS-R	XLS-R + KenLM 3-gram	Whisper
Single speaker, quiet environment.	idem eyen esie isọ̃ñke, enye odu ke ufọk ibọk.	idem eyen esisọ̃ñke enye odu ke ufọk ibọk	idem eyen esie isọ̃ñke enye odu ke ufọk ibọk	idem eyen esi isọ̃ñke, enye eduk ufọk ibọk.
Multi-speaker, noisy environment.	ndien akpa ñke nyenede edi, oyom ikpọ ñkpọ.	nyenakpa ke eyen ededi ikaa oyom ikpọ̃ñ kpọ	ndian akpañ ke enyenede edi yak oyom ikpọ ñkpọ	ndien akpañke enyenede edi, oyom ikpọ̃ñkpọ.
Multi-speaker, noisy environment.	ọkọ̃ñọ mbañ enañ. Utu ke mbọk osun udi. Yak edim, o, edep.	añ ambae namikpokke mbọk osim di ia edim aaedem	ọkọ̃ñọ mbañ enam koko ke mbọk esin udi yak edim edem	Okoñ amaña enañ. Otuk embọk ọsunu di. Ya edim, o, edep.

Table 3: Qualitative semantic comparison of XLS-R, XLS-R + KenLM, and Whisper on Efik ASR under domain and speaker mismatch

ChrF. This approach approximates downstream usability, reflecting scenarios where ASR outputs are consumed by non-Efik speakers or integrated into multilingual pipelines.

As shown in Table 4, XLS-R + KenLM-3-gram achieves the highest BLEU and ChrF scores across both single-speaker and multi-speaker conditions, consistently outperforming both the base XLS-R model and Whisper. Notably, under noisy multi-speaker conditions where Whisper previously showed an advantage in ChrF over the base XLS-R model, the KenLM-enhanced XLS-R surpasses Whisper across all metrics. These results demonstrate that incorporating a language model not only improves phonetic accuracy (as reflected in WER/CER) but also yields substantial gains in semantic preservation, even under challenging acoustic conditions. This underscores the value of leveraging language models for low-resource tonal languages like Efik, where both phonetic and semantic fidelity are critical for downstream usability.

6.3. Discussion

Despite being trained on limited audio data from a single speaker in a quiet environment, both models demonstrate notable gains in Efik automatic speech recognition, highlighting the effectiveness of modern self-supervised and encoder–decoder architectures in extremely low-resource tonal settings.

On a held-out single-speaker evaluation set recorded independently from the training data, the fine-tuned XLS-R-300M model achieved a WER of 29.2% and a CER of 6.4%, outperforming a Whisper baseline trained on the same Efik corpus. The addition of a 3-gram KenLM language model further improved performance significantly, achieving a WER of 10.86% and a CER of 3.16%, a relative reduction of 62.8% in WER and 50.6% in CER over the base XLS-R model. This substantial gain underscores the value of incorporating language model priors, particularly for low-resource tonal languages where phonetic ambiguity is high.

Under real-world noisy and multi-speaker conditions, Whisper exhibits greater robustness and semantic fidelity compared to the base XLS-R model, likely due to its integrated autoregressive language modeling and encoder-decoder architecture, despite Efik's morphological and tonal complexity. However, the KenLM-enhanced XLS-R model bridges this gap considerably, demonstrating improved semantic preservation in translation-based evaluations while maintaining its phonetic accuracy advantages. These findings suggest that effective language model integration, whether through an external n-gram LM or an architecture-internal component, is critical for achieving robust ASR performance in low-resource tonal languages under domain shift.

7. Future Work

Our results indicate that SOTA ASR models can be significantly improved for Efik with more high-quality, multi-speaker, and well-annotated speech data. We plan to curate additional audio spanning multiple domains, speakers ranging from children to adults, and diverse noisy environments. Expanding the dataset beyond the current 3 hours of single-speaker recordings will help the models better capture tonal variations, semantic nuances, and morphological complexity inherent to Efik. Additionally, we will explore data augmentation techniques such as speed and pitch perturbation and noise injection to further enhance robustness and generalization.

8. Conclusion

We evaluated the adaptation of two pre-trained SOTA ASR models, XLS-R and Whisper, for Efik, a tonal language with morphological complexity.

Scenario	Model	BLEU	ChrF	Human Score	Notes
Single speaker, quiet environment.	XLS-R	43.01	74.27	8/10	Very close to reference.
Single speaker, quiet environment.	Whisper	14.21	35.06	7/10	Slight token error, meaning still intact.
Single speaker, quiet environment.	**XLS-R + KenLM**	**59.54**	**80.40**	**9.2/10**	**Very close to reference with minimal token errors**
Multi-speaker, noisy environment.	XLS-R	4.07	13.59	2.8/10	Errors due to unseen speakers and noisy background. Severe degradation in semantic coherence is observed.
Multi-speaker, noisy environment.	Whisper	4.52	28.82	3.6/10	Slight errors in unseen speakers and noisy background, slight semantic preservation and meaning.
Multi-speaker, noisy environment.	**XLS-R + KenLM**	**6.88**	**29.95**	**3.9/10**	**Moderate improvement over base XLS-R; better character-level preservation than Whisper, though semantic coherence remains challenged by domain shift.**

Table 4: Qualitative semantic comparison of XLS-R and Whisper on Efik ASR under domain and speaker mismatch.

Both models were fine-tuned on 3 hours of single-speaker audio (2,632 clips) over 12 epochs each, showing significant improvements in low-resource settings.

The integration of a 3-gram KenLM language model with XLS-R yielded substantial gains across all evaluation conditions. KenLM-enhanced XLS-R achieved the lowest overall WER (10.86%) and CER (3.16%) on the single-speaker test set, outperforming both the base XLS-R model and Whisper. Under noisy multi-speaker conditions, the KenLM-enhanced model also demonstrated improved robustness compared to the base XLS-R model, achieving competitive performance with Whisper while maintaining its phonetic accuracy advantages.

Whisper's ability to model punctuation contributed to better context preservation in certain cases, highlighting the importance of architectural choices in low-resource ASR. Our results demonstrate that combining self-supervised acoustic models with language model integration offers a promising pathway for building robust ASR systems for under-resourced tonal languages.

9. Ethics Statement

We obtained informed consent from all volunteers who participated in the data recording process. The dataset does not contain sensitive, personal, or otherwise violatory content. All recordings were collected and used in accordance with ethical research practices.

10. Acknowledgements

We would like to sincerely thank all the volunteers, native speakers, and linguists who contributed to the recording process and assisted in the evaluation of the results.

11. Bibliographical References

Jade Abbott and Laura Martinus. 2019. Benchmarking neural machine translation for southern african languages. In *Proceedings of the Workshop on Widening NLP*, pages 98–101, Florence, Italy. Association for Computational Linguistics.

David Ifeoluwa Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba O. Alabi, Shamsuddeen H. Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Elvis Mboning, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo L. Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Adeyemi, Gilles Q. Hacheme, Idris Abdulmu-

min, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu Ngoli, and Dietrich Klakow. 2022. Masakhaner 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint*.

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

Yonas Chanie, Moayad Elamin, Paul Ewuzie, and Samuel Rutunda. 2023. Multilingual automatic speech recognition for kinyarwanda, swahili, and luganda. In *Conference Proceedings*.

Brian DeRenzi, Anna Dixon, Mohamed Aymane Farhi, and Christian Resch. 2025. Synthetic voice data for automatic speech recognition in african languages. *arXiv preprint*.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197. Association for Computational Linguistics.

Sukairaj Hafiz Imam, Babangida Sani, Dawit Ketema Gete, Bedru Yimam Ahamed, Ibrahim Said Ahmad, Idris Abdulmumin, Seid Muhie Yimam, Muhammad Yahuza Bello, and Shamsuddeen Hassan Muhammad. 2025. Automatic speech recognition for african low-resource languages: Challenges and future directions. In *Proceedings of the Workshop on African NLP*.

Thennal D K, Jesin James, Deepa P Gopinath, and Muhammed Ashraf K. 2024. Advocating character error rate for multilingual asr evaluation. *arXiv preprint*.

Eyo Mensah and Eyamba Mensah. 2014. The adaptation of english consonants by efik learners of english. *English Language Teaching*, 7(3).

Alvin Nahabwe, Sulaiman Kagumire, Denis Musinguzi, Bruno Beijuka, Jonah Mubuuke Kyagaba, Peter Nabende, Andrew Katumba, and Joyce Nakatumba-Nabende. 2025. Benchmarking automatic speech recognition models for african languages. *arXiv preprint*.

Offiong Ani Offiong and Stella Ansa. 2013. The efik language: A historical profile. *Research in Humanities and Social Sciences*, 3(6).

Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, et al. 2023. Afrispeech-200: Pan-african accented speech dataset for clinical and general domain asr. *Transactions of the Association for Computational Linguistics*, 11:1669–1685.

Youngja Park, Siddharth Patwardhan, Karthik Visweswariah, and Stephen C. Gates. 2008. An empirical analysis of word error rate and keyword error rate. In *INTERSPEECH 2008*.

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

Amina Mardiyyah Rufai, Afolabi Abeeb, Esther Oduntan, Tayo Arulogun, Oluwabukola Adegboro, and Daniel Ajisafe. 2020. Towards end-to-end training of automatic speech recognition for nigerian pidgin. *arXiv preprint*.

Anuoluwapo Aremu Tolúlópè Ógúnrèmí, Kòlá Túbòsún. 2024. Ìròyìnspeech: A multi-purpose yorùbá speech corpus. In *Proceedings of LREC 2024*.

Haihua Xu, Hang Su, Chongjia Ni, Xiong Xiao, Hao Huang, Eng-Siong Chng, and Haizhou Li. 2016. Semi-supervised and cross-lingual knowledge transfer learnings for dnn hybrid acoustic models under low-resource conditions. In *INTERSPEECH 2016*.