Benchmarking Self-Supervised Speech Models on Multilingual Nigerian Speech

Omotayo Omoyemi
University of Derby, UK
etayo25@gmail.com

Ifeoluwa Oladeni
National Open University of Nigeria
oladeniifeoluwa123@gmail.com

Abstract

Self-supervised speech models such as Whisper and wav2vec 2.0 have significantly advanced automatic speech recognition (ASR) performance for high-resource languages. However, their robustness and generalization to underrepresented African languages remain insufficiently studied.

In this work, we present a systematic benchmark of modern self-supervised ASR models on a multilingual Nigerian speech corpus comprising English, Hausa, Igbo, and Yoruba. Using the Nigerian Common Voice dataset (158 hours), we evaluate zero-shot performance of pretrained models and compare it with supervised adaptation using fine-tuning of multilingual speech encoders. We report Word Error Rate (WER) and Character Error Rate (CER) across languages and analyze the effect of supervised adaptation and cross-language transfer.

Our results show that zero-shot ASR performance is substantially degraded for Nigerian languages compared to widely represented benchmark languages. Supervised fine-tuning consistently improves recognition accuracy, although the magnitude of improvement varies across languages and depends on the compatibility between the pretrained checkpoint and the target language. In particular, adaptation from a Hausa-pretrained XLS-R model yields strong gains for Hausa but more limited improvements for Igbo, highlighting the importance of language-specific training data.

These findings demonstrate that multilingual pretraining alone is insufficient for reliable ASR in underrepresented African languages and that supervised adaptation remains necessary for robust deployment. The study provides reproducible benchmarks for multilingual ASR evaluation in African contexts and offers practical guidance for adapting large-scale speech models to underrepresented languages.

1 Introduction

Self-supervised and weakly supervised pretraining has become a dominant paradigm in automatic speech recognition (ASR), enabling strong performance with limited labeled data and improving robustness across domains. Representative examples include wav2vec 2.0, which learns speech representations from raw audio and achieves competitive word error rates after fine-tuning [Baevski et al., 2020], and Whisper, which scales weakly supervised training to hundreds of thousands of hours of web data and demonstrates broad generalization across benchmarks [Radford et al., 2022]. Cross-lingual pretraining further extends these gains to multilingual settings, as shown by large-scale wav2vec-style models trained on hundreds of thousands of hours spanning many languages [Babu et al., 2021; Conneau et al., 2021].

Despite this progress, recent work shows that generalization remains uneven for underrepresented languages, accents, and demographic groups. Performance disparities in ASR are often linked to imbalanced data coverage and limited linguistic diversity, with surveys highlighting persistent challenges for underrepresented African languages [Nakatumba Nabende et al., 2025]. In addition, accent and language labels in widely used speech datasets are often self-reported and inconsistently defined, which can complicate both training and evaluation [Reid and Williams, 2023]. These issues motivate systematic benchmarking to determine where modern pretrained ASR models succeed, where they fail, and how much adaptation is required for reliable deployment.

Nigeria provides a particularly relevant testbed for this analysis. It is linguistically diverse and includes several widely spoken languages, such as English, Hausa, Igbo, and Yoruba, that remain underrepresented in standard ASR benchmarks. At the same time, recent work has begun to release African-accented speech resources, including AfriSpeech-200 for Pan-African accented English [Olatunji et al., 2023] and Afrispeech-Dialog for conversational African-accented English [Sanni et al., 2025], highlighting the need for broader multilingual evaluations that include multiple African languages rather than English alone.

In this work, we benchmark modern pretrained ASR models on a multilingual Nigerian speech corpus with official train/dev/test splits. We evaluate (i) zero-shot recognition using a large pretrained model, and (ii) supervised adaptation using fine-tuning of an XLS-R checkpoint on individual languages. We report Word Error Rate (WER) and Character Error Rate (CER) across languages and compare performance

under zero-shot and fine-tuned settings. Our results show that zero-shot multi-lingual ASR performs poorly for Nigerian languages, while supervised adaptation substantially reduces error rates, although the magnitude of improvement varies across languages. This study provides empirical evidence on the limits of multilingual pretraining and the importance of language-specific adaptation for underrepresented African ASR, using reproducible evaluation protocols based on widely used dataset tooling [Lhoest et al., 2021] and multilingual speech corpus design practices inspired by Common Voice [Ardila et al., 2020].

2 Related Work

2.1 Multilingual and Large-Scale ASR Benchmarks

The evaluation of multilingual ASR systems has expanded beyond high-resource settings toward more inclusive and globally representative benchmarks. The ML-SUPERB 2.0 Challenge introduces a large-scale multilingual evaluation suite covering over 200 language varieties and demonstrates that even state-of-the-art pretrained models exhibit significant performance variability across languages and dialects (Chen et al., 2025). These findings reinforce that zero-shot robustness cannot be assumed for underrepresented speech communities.

Complementing this direction, recent industrial research has released African-focused ASR benchmarks such as PazaBench, which provides standardized evaluation resources for 39 African languages and explicitly targets deployment in underrepresented settings (Muchai et al., 2026). These efforts collectively highlight the need for systematic, reproducible benchmarking of pretrained speech models in linguistically diverse African contexts.

2.2 Efficient Adaptation for Low-Resource Languages

Beyond benchmarking, methodological work has examined how to adapt large speech models to underrepresented languages efficiently. Parameter-efficient fine-tuning approaches, including adapter-based methods, have demonstrated competitive performance in extremely underrepresented scenarios without full model retraining (Mainzinger, 2024). Similarly, efficient ASR training frameworks tailored for underrepresented languages have been proposed to reduce data and computational requirements while preserving accuracy (Bandarupalli, 2025).

These studies suggest that adaptation strategy plays a critical role in determining whether large pretrained models can meaningfully generalize to underrepresented African languages.

2.3 Accent Robustness and Fairness in ASR

A growing body of work has examined disparities in ASR performance across accents and demographic groups. Accent-invariant modeling approaches, such as saliency-driven spectrogram masking, have been proposed to enhance robustness to dialectal variation and reduce word error rate gaps between accent groups (Sameti et al., 2025). In parallel, fairness-aware fine-tuning techniques have shown that multilingual ASR models can be optimized to mitigate performance disparities across demographic attributes without substantial degradation in overall accuracy (Swain et al., 2024).

Systematic reviews focused specifically on African underrepresented languages further document persistent gaps in dataset coverage, evaluation consistency, and bias mitigation strategies, emphasizing the importance of transparent benchmarking practices (Imam et al., 2025).

2.4 Implications for African Multilingual ASR

Collectively, recent benchmarking initiatives, adaptation techniques, and fairness-aware modeling strategies underscore a central insight: while large pretrained speech models have dramatically advanced ASR, their performance in multilingual African contexts remains uneven and insufficiently characterized. The literature increasingly calls for reproducible, language-specific evaluation frameworks grounded in realistic deployment settings (Chen et al., 2025; Muchai et al., 2026).

In this work, we respond to this gap by conducting a systematic evaluation of modern pretrained ASR models on multilingual Nigerian speech, analyzing zero-shot generalization, supervised adaptation, and cross-language transfer across four languages.

3 Dataset

Language	#Utterances	Hours	Mean Dur (s)	Median Dur (s)	P95 Dur (s)	Mean Chars
English	3,402	11.08	11.72	6.7	39.0	127.3
Hausa	9,008	10.76	4.30	4.1	6.7	68.4
Igbo	5,714	8.60	5.42	5.2	9.1	45.8
Yoruba	4,171	6.82	5.88	5.8	8.3	57.6

Table 1: Summary Statistics of the Nigerian Multilingual Speech Corpus (All Splits Combined)

We conduct our experiments using a multilingual Nigerian speech corpus comprising English, Hausa, Igbo, and Yoruba. The dataset provides official train, validation, and test splits for each language. Statistics reported in Table 1 are computed across all splits combined.

The corpus contains a total of approximately 37.25 hours of transcribed speech distributed unevenly across languages. Hausa con-

tains the largest number of utterances (9,008), while Yoruba contains the smallest total duration (6.82 hours). English exhibits notably longer utterances on average (mean duration 11.72 seconds) compared to the other languages (4–6 seconds). The 95th percentile duration for English (39.0 seconds) is substantially higher than for Hausa, Igbo, and Yoruba (all below 10 seconds), indicating structural differences in utterance length distributions across languages.
All audio recordings were resampled to 16 kHz to ensure compatibility with pretrained ASR models and to maintain consistent acoustic representation across languages. Following the dataset documentation, transcripts were normalized by removing surrounding quotation marks and ensuring sentence-final punctuation. No additional text cleaning or filtering was performed.
Table 1 also reports transcript length statistics. English transcripts are considerably longer on average (127.3 characters) than the other languages, with Igbo exhibiting the shortest mean transcript length (45.8 characters). These differences in duration and transcript length are expected to influence recognition difficulty and are considered in the interpretation of experimental results.
The dataset version used in this study was frozen after preprocessing to ensure reproducibility.

4 Method

4.1 Model Selection

We evaluate automatic speech recognition performance using Whisper-small as a zero-shot baseline. Whisper is a large-scale encoder–decoder model trained on multilingual speech–text data using weak supervision. We select the "small" variant to balance computational feasibility on CPU hardware and model capacity. For subsequent supervised experiments, we fine-tune multilingual speech encoders (e.g., XLS-R) using a CTC objective.

4.2 Zero-Shot Inference Protocol

Zero-shot evaluation is performed using the faster-whisper implementation with int8 CPU inference. Audio inputs are resampled to 16 kHz as described in Section 3. Decoding is performed with voice activity detection enabled.
Where supported by the tokenizer, forced language decoding is applied (English: "en", Hausa: "ha", Yoruba: "yo"). For Igbo, language auto-detection is used because the faster-whisper tokenizer does not include an Igbo language code. Transcripts are normalized using the same preprocessing described in Section 3 prior to computing evaluation metrics.

4.3 Fine-Tuning Strategy

For supervised adaptation, we fine-tune a multilingual speech encoder using the Connectionist Temporal Classification (CTC) objective. Experiments are conducted using the pretrained XLS-R checkpoint Mofe/xls-r-hausa-40, which was originally trained for Hausa speech recognition. This checkpoint is used as the starting point for all adaptation experiments to study both language-specific fine-tuning and cross-language transfer.
Fine-tuning is performed separately for each language using the official training split, with model selection based on validation performance. In particular, we fine-tune the Hausa checkpoint on the Hausa subset and additionally adapt the same checkpoint to Igbo to evaluate cross-language adaptation behavior.
Training is performed using AdamW with a linear learning-rate schedule. Because experiments are conducted on CPU hardware, we use small batch sizes together with gradient accumulation to ensure stable optimization. The feature encoder of the pretrained model is frozen during fine-tuning to reduce memory usage and improve convergence stability.
To avoid invalid CTC alignments, examples where the label sequence is longer than half the number of input frames are filtered prior to training. This filtering rule ensures that the CTC loss remains well-defined and prevents training instabilities.
All experiments use fixed train/validation/test splits to ensure comparability across languages and training regimes.

4.4 Evaluation Metrics

Performance is measured using Word Error Rate (WER) and Character Error Rate (CER), computed on the official test split. WER captures token-level transcription accuracy, while CER provides robustness for morphologically rich or orthographically variable languages. Metrics are computed after transcript normalization to ensure consistency across languages.

4.5 Implementation Details

All experiments are implemented in Python using the Hugging Face Transformers and Datasets libraries, together with faster-whisper for zero-shot inference. Zero-shot evaluation is conducted on a CPU-only system (Intel i7-10510U, 16 GB RAM) using int8 quantization. Fine-tuning experiments are also performed on CPU hardware using gradient accumulation to simulate larger batch sizes.
Dataset preprocessing, resampling to 16 kHz, manifest generation, and filtering of invalid CTC alignments are fixed prior to experimentation to ensure reproducibility. All runs use the same official train/validation/test splits provided with the corpus.

5 Experiments

5.1 Zero-Shot Evaluation

We first evaluate Whisper-small in a zero-shot setting on the official test split for each language. No supervised adaptation is performed. Audio inputs are resampled to 16 kHz and decoded using faster-whisper with int8 CPU inference, as described in Section 4.
Table 2 reports the average Word Error Rate (WER) and Character Error Rate (CER) for English, Hausa, Igbo, and Yoruba. Forced language decoding is applied for English, Hausa, and Yoruba. For Igbo, language auto-detection is used because the faster-whisper tokenizer does not include an Igbo language code.

Language	Avg WER	Avg CER	Test Utterances	Decoding Language Setting
English	0.2739	0.1500	341	Forced (en)
Hausa	0.9546	0.3604	901	Forced (ha)
Igbo	1.4532	0.9965	572	Auto-detect (no 'ig' code)
Yoruba	1.5025	1.1076	418	Forced (yo)

Table 2: Zero-shot ASR performance of Whisper-small on the Nigerian multilingual test split.

Note: WER/CER are computed on the official test split after transcript normalization and 16 kHz resampling. For Igbo, faster-whisper does not provide an "ig" language code, so decoding used language auto-detection.

Zero-shot performance varies substantially across languages. English achieves comparatively low error rates, while Hausa exhibits significantly higher WER despite forced language decoding. Performance on Igbo and Yoruba is extremely poor, with WER exceeding 1.45 and CER approaching or exceeding 1.0. These results indicate that Whisper-small generalizes unevenly to Nigerian languages, particularly for languages that are likely underrepresented in the model's pretraining data.

The sharp degradation for Igbo and Yoruba suggests that multilingual pretraining alone is insufficient to ensure robust recognition for underrepresented African languages. These findings motivate the supervised fine-tuning experiments described in the next subsection.

5.2 Supervised Fine-Tuning (Hausa)

To examine whether supervised adaptation can mitigate the performance gap observed in the zero-shot setting, we fine-tune a pretrained Hausa ASR checkpoint based on XLS-R using the Hausa portion of the Nigerian multilingual speech corpus. Training is performed on the official training split, model selection is based on validation performance, and final evaluation is conducted on the held-out test split.

Table 3 summarizes the results. After one epoch of fine-tuning, the validation word error rate (WER) reaches 0.3563, with a character error rate (CER) of 0.0912. After the second epoch, the validation WER improves slightly to 0.3485 with a CER of 0.0888, indicating stable convergence during training.

On the Hausa test split, the fine-tuned model achieves a WER of 0.5306 and a CER of 0.5268. Compared with the zero-shot Whisper-small baseline (WER 0.9546), supervised fine-tuning reduces the WER by 44.4% relative. This substantial improvement demonstrates that multilingual pretrained speech models benefit strongly from language-specific adaptation when applied to underrepresented African languages.

Although the fine-tuned model does not reach the performance typically observed for high-resource languages, the results confirm that zero-shot transfer alone is insufficient for reliable recognition in Hausa, while supervised fine-tuning significantly improves recognition accuracy. This finding supports the hypothesis that underrepresented languages require targeted adaptation even when using large multilingual pretrained models.

Model / Setting	Split	WER	CER	Notes
Whisper-small (zero-shot)	Test	0.9546	0.3604	No supervised adaptation
XLS-R Hausa (fine-tuned)	Validation	0.3563	0.0912	Epoch 1
XLS-R Hausa (fine-tuned)	Validation	0.3485	0.0888	Epoch 2
XLS-R Hausa (fine-tuned)	Test	0.5306	0.5268	Verified evaluation

Table 3: Supervised fine-tuning results for Hausa ASR. The adapted XLS-R model significantly outperforms the zero-shot Whisper-small baseline on the Hausa test split

5.3 Cross Language Adaptation (Igbo)

To examine whether the gains observed for Hausa generalize to another Nigerian language, we further adapt the pretrained XLS-R Hausa checkpoint on the Igbo portion of the corpus. Table 4 reports the validation and test results. After one epoch of fine-tuning, the validation WER reaches 0.9597, improving slightly to 0.9463 after the second epoch. On the Igbo test split, the adapted model achieves a WER of 0.9464 and a CER of 0.4341.

Model / Setting	Split	WER	CER	Notes
Whisper-small (zero-shot)	Test	1.4532	0.9965	No supervised adaptation
XLS-R Hausa checkpoint adapted to Igbo	Validation	0.9597	0.4584	Epoch 1
XLS-R Hausa checkpoint adapted to Igbo	Validation	0.9463	0.4385	Epoch 2
XLS-R Hausa checkpoint adapted to Igbo	Test	0.9464	0.4341	Final test evaluation

Table 4: Supervised fine-tuning results for Igbo ASR. Cross-language adaptation from a Hausa XLS-R checkpoint improves performance over the zero-shot baseline on the Igbo test split.

Compared with the zero-shot Whisper-small baseline (WER 1.4532), supervised adaptation reduces the Igbo test WER by 34.9% relative. Although the gains are smaller than those observed for Hausa, the results still indicate that language-specific or cross-language supervised adaptation improves recognition performance substantially over zero-shot transfer alone. This suggests that pretrained multilingual speech models can benefit from further adaptation even when the starting checkpoint was optimized for a different African language.

6. Discussion

The experimental results highlight several important observations about multilingual speech recognition for underrepresented Nigerian languages. First, the zero-shot evaluation demonstrates that performance varies substantially across languages, even when using the same large pretrained model. While Whisper-small performs reasonably well for English, recognition accuracy drops sharply for Hausa, Igbo, and Yoruba, with error rates exceeding those typically observed for high-resource languages. This suggests that multilingual pretraining alone is insufficient to ensure reliable performance for underrepresented languages.

Second, supervised fine-tuning consistently improves recognition accuracy across languages. For Hausa, adapting an XLS-R Hausa checkpoint on the Nigerian Hausa corpus reduces the test WER from 0.9546 in the zero-shot setting to 0.5306 after fine-tuning, corresponding to a relative reduction of 44.4%. A similar trend is observed for Igbo, where adapting the same checkpoint reduces the test WER from 1.4532 to 0.9464, a relative reduction of 34.9%. These results confirm that language-specific training data remains crucial even when starting from large multilingual pretrained models.

However, the magnitude of improvement differs between languages. The gains obtained for Hausa are larger than those for Igbo, which may be explained by differences in training data size, phonetic characteristics, or the mismatch between the pretrained checkpoint and the target language. In this study, the adaptation experiments use a checkpoint originally trained for Hausa, which may provide a better initialization for Hausa than for Igbo. This indicates that cross-language transfer is possible but not uniform, and the effectiveness of adaptation depends on both the similarity between languages and the amount of available training data.

Overall, the results suggest that multilingual speech models provide a strong starting point, but reliable recognition for underrepresented African languages still requires targeted adaptation. Zero-shot transfer alone is not sufficient, and supervised fine-tuning can substantially reduce error rates, although performance remains below that typically achieved for high-resource languages. These findings highlight the need for more language-specific resources and better multilingual training strategies to support speech recognition in underrepresented languages.

This study has several limitations. First, experiments were conducted on CPU hardware, which constrained the size of models and the number of adaptations runs that could be performed. Second, adaptation experiments were limited to a single pretrained checkpoint, and results may differ when using other multilingual models. Third, the dataset size for some languages is relatively small, which may affect the stability of training and evaluation. Future work should investigate larger multilingual corpora, additional pretrained checkpoints, and parameter-efficient adaptation methods to further improve ASR performance for African languages.

7 Conclusion

This paper investigated multilingual automatic speech recognition for Nigerian languages using zero-shot inference and supervised fine-tuning with pretrained speech models. Experiments were conducted on a Nigerian multilingual speech corpus containing English, Hausa, Igbo, and Yoruba. In the zero-shot setting, Whisper-small achieved reasonable performance for English but produced high error rates for the Nigerian languages, with test WER values exceeding 0.95 for Hausa and above 1.4 for Igbo. These results indicate that multilingual pretraining alone does not guarantee reliable recognition for underrepresented languages.

To address this limitation, we evaluated supervised adaptation using an XLS-R checkpoint. Fine-tuning on the Hausa subset reduced the test WER from 0.9546 in the zero-shot setting to 0.5306, corresponding to a relative improvement of 44.4%. A similar experiment on Igbo reduced the test WER from 1.4532 to 0.9464, yielding a relative improvement of 34.9%. These results show that supervised fine-tuning consistently improves recognition accuracy, although the magnitude of improvement varies across languages.

The difference in improvement between Hausa and Igbo suggests that cross-language transfer from a pretrained checkpoint is not uniform. In this study, the starting checkpoint was trained for Hausa, which likely provided a better initialization for Hausa than for Igbo. This highlights the importance of language-specific training data and indicates that multilingual speech models still require targeted adaptation for underrepresented languages.

Overall, the experiments demonstrate that zero-shot multilingual ASR remains insufficient for reliable recognition of Nigerian languages, but supervised fine-tuning can substantially reduce error rates. Future work should explore larger multilingual training sets, language-balanced pretraining, and adaptation methods designed specifically for underrepresented African languages.

References

[Ardila et al., 2020] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common Voice: A Massively-Multilingual Speech Corpus. In Proceedings of LREC, 2020.

[Baevski et al., 2020] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In Advances in Neural Information Processing Systems (NeurIPS), 2020.

[Babu et al., 2021] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, and Alexei Baevski. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. arXiv preprint, 2021.

[Conneau et al., 2021] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In Proceedings of Interspeech, 2021.

[Lhoest et al., 2021] Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander M. Rush, and Thomas Wolf. Datasets: A Community Library for Natural Language Processing. arXiv preprint, 2021.

[Nakatumba Nabende et al., 2025] Joyce Nakatumba Nabende, Sulaiman Kagumire, Caroline Kantono, and Peter Nabende. A Systematic Literature Review on Bias Evaluation and Mitigation in Automatic Speech Recognition Models for Low-Resource African Languages. ACM Computing Surveys, 58(4), 2025.

[Olatunji et al., 2023] Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, Sahib Singh, Bonaventure F. P. Dossou, Joanne Osuchukwu, Salomey Osei, Atnafu Lambebo Tonja, Naome Etori, and Clinton Mbataku. AfriSpeech-200: Pan-African Accented Speech Dataset for Clinical and General Domain ASR. Transactions of the Association for Computational Linguistics, 2023.

[Radford et al., 2022] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision. arXiv preprint, 2022.

[Reid and Williams, 2023] Kathy Reid and Dominic Williams. Common Voice and Accent Choice: Data Contributors Self-Describe Their Spoken Accents in Diverse Ways. In Proceedings of ACM Conference, 2023.

[Sanni et al., 2025] Mardhiyah Sanni, Tassallah Abdullahi, Devendra D. Kayande, Emmanuel Ayodele, Naome A. Etori, Michael S. Mollel, Moshood Yekini, Chibuzor Okocha, Lukman E. Ismaila, Folafunmi Omofoye, Boluwatife A. Adewale, and Tobi Olatunji. Afrispeech-Dialog: A Benchmark Dataset for Spontaneous English Conversations in Healthcare and Beyond. In Proceedings of NAACL, 2025.

[Chen et al., 2025] William Chen, Chutong Meng, et al. The ML-SUPERB 2.0 Challenge: Towards Inclusive ASR Benchmarking for All Language Varieties. arXiv preprint, 2025.

[Muchai et al., 2026] Mercy Muchai, Kevin Chege, et al. Paza: Introducing Automatic Speech Recognition Benchmarks and Models for Low Resource Languages. Microsoft Research, 2026.

[Bandarupalli, 2025] S. Bandarupalli. Efficient ASR for Low-Resource Languages. arXiv preprint, 2025.

[Sameti et al., 2025] Mohammad Hossein Sameti, Sepehr Harfi Moridani, Ali Zarean, and Hossein Sameti. Accent-Invariant Automatic Speech Recognition via Saliency-Driven Spectrogram Masking. arXiv preprint, 2025.

[Swain et al., 2024] M. Swain, et al. On Mitigating Performance Disparities in Multilingual ASR via Fairness-Aware Fine-Tuning. In Proceedings of EMNLP, 2024.

[Imam et al., 2025] Sukairaj Hafiz Imam, et al. Automatic Speech Recognition for African Low-Resource Languages: A Systematic Literature Review. ACM Computing Surveys, 2025.

[Mainzinger, 2024] J. Mainzinger. Fine-Tuning ASR Models for Very Low-Resource Languages. In Proceedings of ACL Student Research Workshop, 2024.