

PazaBench: Large-Scale ASR Benchmark for Low-Resource Languages in Africa

Mercy Muchai, Nick Mumbo Mwangi, Kevin Chege, Samuel Chege Maina

Microsoft Research Africa, Nairobi

{mercymuchai, v-nickmwangi, v-kevinchege, samuelmaina}@microsoft.com

Abstract

Low-resource languages remain underrepresented in ASR benchmarks, limiting the ability to reliably assess and compare model performance. We present a large-scale benchmark for automatic speech recognition (ASR) on low-resource languages in Africa. The benchmark harmonizes six public corpora into a unified, reproducible evaluation framework and evaluates 52 state-of-the-art ASR models spanning three model architectural paradigms across 39 African languages on diverse datasets. Models are evaluated using standard accuracy and efficiency metrics, word error rate (WER), character error rate (CER), and inverse real-time factor (RTFx). This benchmark introduces architecture-aware comparison that systematically analyzes both individual models and broader model families. In addition, we quantify linguistic variance across language families, revealing how these factors shape ASR performance across diverse African languages. We also provide a comprehensive accuracy and efficiency trade-off analysis directly relevant to deployment in low-resource settings. Together, the benchmark, evaluation pipeline, and results support transparent, reproducible, and deployment-aware evaluation of ASR models for low-resource languages.

1 Introduction

Recent progress in automatic speech recognition (ASR) has been driven by large-scale multilingual training, self-supervised representation learning, and the integration of speech with large language models, resulting in strong performance on widely used benchmarks for high-resource languages [Radford *et al.*, 2023]. These advances have broadened the ASR landscape, spanning encoder-only and encoder-decoder architectures as well as emerging LLM-based models, while substantially improving cross-lingual transfer capabilities.

However, evaluation practices have not kept pace in low-resource settings: Languages with limited labeled data introduce challenges related to linguistic diversity, morphology,

orthography, and deployment constraints that are not well reflected in the existing benchmarks. Many multilingual ASR evaluations focus on aggregate results or high-resource subsets, which limits insights into model behavior in data-scarce and/or deployment-constrained settings.

African languages illustrate this gap clearly. Despite their linguistic diversity and societal relevance, they remain underrepresented in many multilingual ASR benchmarks, and when included, results are often summarized using macro-averaged metrics that obscure language-specific behavior [Pratap *et al.*, 2024]. This not only limits inclusivity but also constrains our ability to capture meaningful variation in model performance, reducing opportunities to uncover insights that such diversity can provide. Prior work has shown that languages with similar data availability can exhibit markedly different performance, reflecting differences in linguistic structure and model inductive biases [Baevski *et al.*, 2020; Conneau *et al.*, 2021]. As a result, existing benchmarks leave critical questions unanswered on ASR performance for African languages.

At the same time, the ASR ecosystem has expanded along architectural lines. Encoder-only models emphasize inference efficiency through non-autoregressive decoding [Baevski *et al.*, 2020; Xu *et al.*, 2023], encoder-decoder models prioritize transcription accuracy and robustness [Radford *et al.*, 2023], and emerging LLM-based approaches integrate ASR into broader multimodal frameworks [Chu *et al.*, 2024; Abouelenin *et al.*, 2025]. Ignoring these architectural differences or inference efficiency on low-resource languages can obscure the roles of model scale, training data, and design choices, reducing the interpretability and deployment relevance of benchmark results.

Contributions. We present PazaBench¹, the first large-scale unified benchmark evaluating 52 publicly available ASR models across 39 low and mid-resource African languages under a standardized evaluation protocol. This benchmark introduces architecture-aware analysis supported by rigorous non-parametric statistical testing. We further quantify linguistic variance and complexity that are substantial enough to render aggregate multilingual averages misleading. Finally, we provide a comprehensive accuracy–efficiency ASR trade-off analysis, offering actionable insights for deploying

¹<https://huggingface.co/spaces/microsoft/paza-bench>

robust ASR systems for low-resource languages.

2 Related Work

Evaluation of ASR models has traditionally relied on standardized benchmark datasets and metrics such as WER and CER, enabling comparison across models and training approaches. These benchmarks have played a central role in tracking ASR progress under ideal conditions.

Several large-scale multilingual datasets support evaluation across languages, including Common Voice [Ardila *et al.*, 2020] and MLS [Pratap *et al.*, 2020]. Additionally, the CMU Wilderness corpus [Black, 2019] provided one of the earliest large-scale multilingual speech collections covering African languages. While these resources provide broad multilingual coverage, current benchmarks predominantly emphasize high-resource languages and offer limited representation of low-resource languages or those with substantial dialectal variation.

Efforts to build speech representations and benchmarks for underrepresented languages have gained momentum. Self-supervised models such as XLS-R [Babu *et al.*, 2022] and cross-lingual wav2vec [Conneau *et al.*, 2021] demonstrated that pretraining on diverse multilingual audio enables transfer to low-resource settings. More targeted efforts include AfriHuBERT [Alabi and others, 2025], a self-supervised model pretrained specifically on African language speech, and mHuBERT-147 [Boito and others, 2024], a compact multilingual speech model covering 147 languages including several African languages. [Chen and others, 2024] further scale self-supervised pretraining to thousands of languages, including many African languages with minimal labelled data. Recent LLM-based systems further expand the scope of ASR evaluation by integrating speech recognition with broader language understanding tasks. Models such as Phi-4 multimodal-instruct [Abouelenin *et al.*, 2025] and Omnilingual ASR [Omnilingual ASR Team *et al.*, 2025] with 1600+ language coverage demonstrate strong performance, placing new demands on benchmarks to support architecture-aware and efficiency-aware evaluation across diverse languages.

Benchmark suites with broader task and model coverage have also emerged. AudioBench [Wang *et al.*, 2025] extends evaluation beyond ASR to include speech translation, audio captioning, and audio question answering, supporting comparison across varying architectures including LLM-based models such as Qwen-Audio [Chu *et al.*, 2024]. However, its ASR evaluation remains largely concentrated on English. The OpenASR leaderboard [Srivastav *et al.*, 2025] introduces standardized evaluation of accuracy and inference efficiency across hybrid, CTC, and transformer-based models, but continues to focus primarily on high-resource languages.

Dedicated benchmarks for non-European language families have also emerged. ML-SUPERB [Shi *et al.*, 2023] provides a standardized multilingual benchmark spanning diverse language families, while IndicSUPERB [Javed and others, 2023] specifically targets Indian languages. For African languages, SimbaBench [Elmadany *et al.*, 2025] aggregates speech data across 61 African languages and provides a leaderboard for ASR and related tasks. However,

these benchmarks do not compare architectural paradigms or jointly evaluate accuracy and inference efficiency.

This highlights a broader evaluation gap: while dedicated African language datasets, models and few benchmarks exist, none jointly evaluate architecture-aware accuracy and efficiency metrics, and language complexity across diverse ASR paradigms for low-resource African languages.

3 Benchmark Construction and Methodology

This benchmark focuses on African languages with heterogeneous scripts, morphology, and data availability, while jointly evaluating transcription accuracy and efficiency under controlled conditions.

3.1 Languages and Dataset Groups

We construct a unified evaluation benchmark covering 39 African languages by aggregating six publicly available corpora: AfriVoices-KE [Wanzare *et al.*, 2026], African Next Voices South Africa [Marivate *et al.*, 2025], ALFFA [Gauthier *et al.*, 2016], Google FLEURS [Conneau *et al.*, 2022], Mozilla Common Voice 23.0 [Ardila *et al.*, 2020], and Naija Voices [Emezue *et al.*, 2025]. These datasets, as shown on Table 4 span diverse domains, ensuring broad acoustic and linguistic coverage. Dataset selection follows five criteria: (i) inclusion of underrepresented African languages, (ii) availability of validated transcriptions, (iii) compatible licensing, (iv) standardized metadata, and (v) existence of held-out evaluation splits. Rather than maximizing raw data volume, the benchmark emphasizes harmonization of heterogeneous resources into a reproducible and comparable evaluation framework.

3.2 Pre-processing and Evaluation Splits

All audio is resampled to 16 kHz mono, with transcripts extracted using dataset-specific alignment rules. Text normalization retains punctuation and speaker metadata for analysis, though neither is used in scoring. Before WER and CER calculation, punctuation is removed, text is case-folded, and whitespace is normalized. For datasets lacking predefined splits (e.g., Naija Voices), data is merged across batches and stratified before sampling with 3% of the data selected for evaluation. The total evaluation dataset sums up to 199,438 utterances across 60 dataset–language splits. Each sample includes a unique ID, audio waveform, transcript, language code, data source, and metadata. Language codes follow the ISO 639-3 standard,² and are provided to models requiring language conditioning using each model’s language code inventory. Languages are further grouped into their corresponding language families based on Ethnologue [Lewis, 2009] to allow metric-level evaluation by language family.

3.3 Data Reliability by Language

To support robust interpretation of the benchmark results, all evaluations are accompanied by sample count by language as shown in Figure 1. This makes languages with limited evaluation data readily identifiable and supports uncertainty-aware

²<https://iso639-3.sil.org>

interpretation of aggregate performance metrics while providing a clearer picture of where additional data availability is most needed.

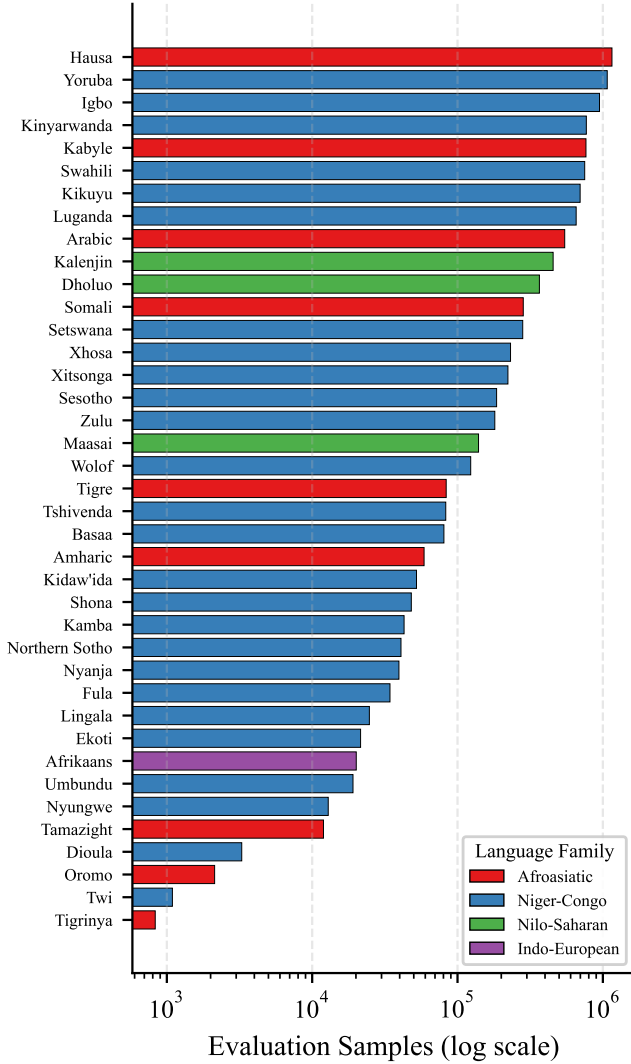


Figure 1: Evaluation sample size per language.

3.4 Benchmark Models

We evaluate 52 publicly available state-of-the-art ASR models spanning 15 model families and three architectural paradigms as shown in Table 8. All models are sourced from open repositories, primarily Hugging Face, ensuring transparency and reproducibility. This diversity allows the benchmark to analyze how model architecture interacts with language coverage, accuracy, and efficiency in low-resource languages.

Although some models are not trained on African languages, all are evaluated under identical cross-lingual conditions, reflecting realistic zero-shot deployment. We confirm that all conclusions remain stable after excluding English-only models, enabling consistent comparison across model

families while highlighting strong candidates for downstream adaptation and finetuning.

3.5 Architectural Paradigms

To enable principled and architecture-aware comparison, we group evaluated models into three high-level architectural classes based on how speech is encoded, how text is generated, and whether decoding is autoregressive or non-autoregressive. This framework enables controlled analysis of systematic trade-offs between accuracy, efficiency, and robustness that are often conflated in prior ASR benchmarks.

Encoder-decoder [Chorowski *et al.*, 2015] ASR models consist of an acoustic encoder paired with an autoregressive text decoder that generates tokens sequentially. This setup enables strong modeling of long-range linguistic dependencies and robust multilingual transfer, but incurs higher inference latency due to sequential decoding.

Encoder-only ASR models employ a deep acoustic encoder with non-autoregressive decoding objectives, eliminating the need for a text decoder. They typically use alignment-based frameworks such as Connectionist Temporal Classification (CTC), Recurrent Neural Network Transducer (RNN-T), or Token-and-Duration Transducer (TDT) [Xu *et al.*, 2023]. This paradigm supports parallel token prediction and low-latency, high-throughput inference, making it suitable for real-time and large-scale deployment.

LLM-based ASR treats speech recognition as one task within a multi-task transformer trained on heterogeneous objectives. Speech is encoded into intermediate representations and transcribed via natural-language generation by a large language model. This enables strong contextual reasoning and instruction-following behavior, but typically at higher computational cost than specialized ASR architectures.

Representative models for each paradigm are summarized in Table 1. This taxonomy separates architectural design choices from model training scale, highlighting trade-offs often obscured in model-level comparisons.

3.6 Inference Setup, Metrics, and Reproducibility

All models are evaluated on a single NVIDIA A100-80GB GPU using official inference settings from their model cards, following the Open ASR Leaderboard evaluation framework [Srivastav *et al.*, 2025]. To ensure fair efficiency comparisons, audio is processed one utterance at a time (batch size of one) with default model decoding settings unless otherwise specified.

We report word error rate (WER), character error rate (CER), and inverse real-time factor (RTFx), where higher RTFx indicates faster-than-real-time inference.

3.7 Statistical Testing Methodology

Shapiro-Wilk tests reject normality for all groups ($p < 10^{-22}$) and Levene’s test confirms unequal variances ($p < 10^{-55}$), ruling out parametric methods. We therefore use the Kruskal-Wallis H -test for all group comparisons, with Dunn’s post-hoc pairwise tests and Holm-Bonferroni correction [Demšar, 2006; Dror *et al.*, 2018]. We report Cliff’s delta [Cliff, 1993] (δ) as a distribution-free pairwise effect

Architecture	Model Families
Encoder-Only	Wav2Vec2 [Baevski <i>et al.</i> , 2020], HuBERT [Hsu <i>et al.</i> , 2021], Data2Vec [Baevski <i>et al.</i> , 2022], MMS [Pratap <i>et al.</i> , 2024], OmniASR-CTC [Omnilingual ASR Team <i>et al.</i> , 2025]
Encoder-Decoder	Whisper [Radford <i>et al.</i> , 2023], Distil-Whisper [Gandhi <i>et al.</i> , 2023], Lite-Whisper [Kamahori <i>et al.</i> , 2025], Moonshine [Jeffries <i>et al.</i> , 2024], Parakeet-TDT [Xu <i>et al.</i> , 2023], SeamlessM4T [Barrault <i>et al.</i> , 2023]
LLM-Based	Phi-4-MM [Abouelenin <i>et al.</i> , 2025], Qwen2-Audio [Chu <i>et al.</i> , 2024], OmniASR-LLM [Omnilingual ASR Team <i>et al.</i> , 2025], Canary [Xu <i>et al.</i> , 2023], Granite [Saon <i>et al.</i> , 2025], Kyutai STT [Zeghidour <i>et al.</i> , 2025]

Table 1: ASR model families grouped by architectural paradigm.

size, and Spearman’s ρ with bootstrap confidence intervals for correlations.

4 Results

We report benchmark results across languages, model families, and architectural paradigms, focusing on understanding systematic trade-offs rather than identifying a single best-performing system.

4.1 Performance by Language

Figure 2 summarizes language-level performance across the 39 evaluated languages. As expected, mid-resource languages (Swahili and Arabic) achieve lower median WER. However, Afroasiatic languages with non-Latin scripts cluster at higher CER, underscoring orthographic challenges for ASR systems trained predominantly on Latin-script languages. Arabic, despite being mid-resource, shows the highest CER (>0.90) alongside Amharic and Tamazight exhibiting extreme variance due to script complexity. Lingala, Luganda, and Nyungwe achieve the lowest median CER (<0.35). These results indicate that script and orthographic complexity can outweigh training data availability as primary drivers of ASR performance under cross-lingual conditions.

4.2 Performance by Model Families

Figure 3 presents aggregated CER performance by model family. The omnilingual ASR family achieves the lowest median accuracy error rates across languages given their high language coverage. Within this family, CTC-based variants achieve higher inference efficiency, while LLM-based variants achieve better accuracy. RTFx varies substantially across families based on architecture. The `nvidia/parakeet-tdt-0.6b-v3` model has the most notable efficiency performance, outperforming all models by

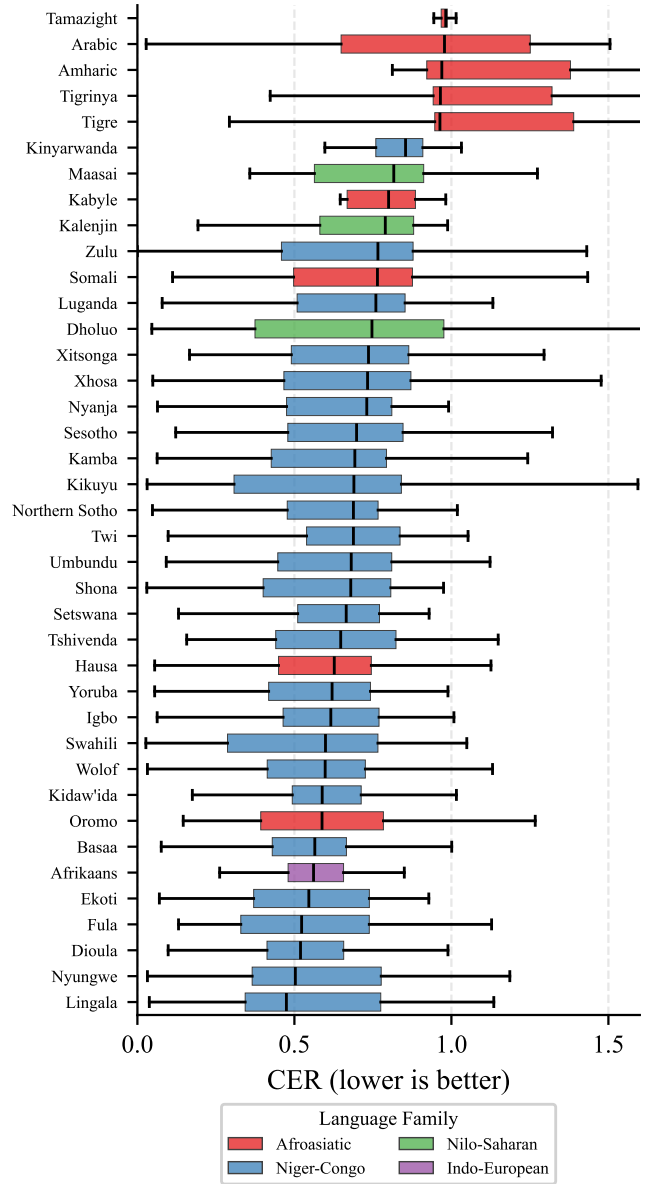


Figure 2: CER distribution across the 39 African languages, aggregated over all models.

more than twice the transcription speed. Table 2 showcases all metrics for the top-performing 15 models.

4.3 Architectural Trade-offs

Figure 4 compares CER distributions across architectural paradigms. A Kruskal-Wallis test confirms that CER distributions differ significantly across the three paradigms ($H = 924.8$) with all pairwise comparisons significant after Holm-Bonferroni correction. Encoder-Decoder vs. Encoder-only shows the largest pairwise separation (Cliff’s $\delta = 0.71$), while Encoder-only vs. LLM-Based yields a negligible effect ($\delta = -0.09$), confirming the comparable accuracy between these two paradigms despite fundamentally different decoding strategies. As shown in Figure 5 RTFx differences

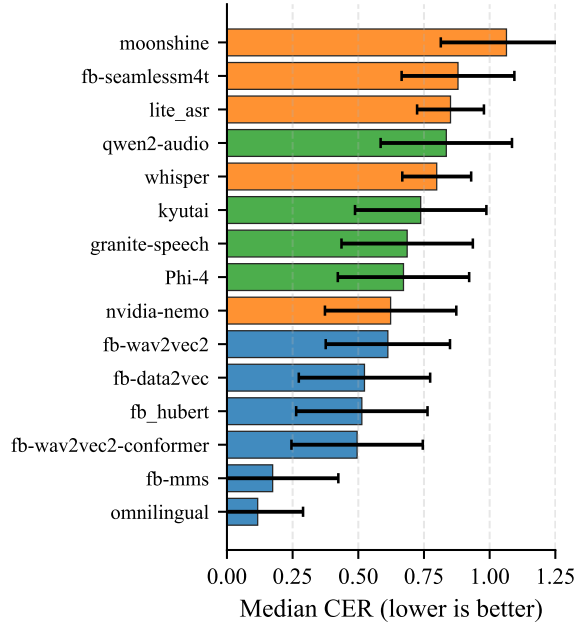


Figure 3: Median CER by model family across all evaluated languages.

Model	Arch.	CER	WER	RTFx
omniASR-LLM-7B	LLM	0.09	0.34	7.5
omniASR-LLM-7B_ZS	LLM	0.10	0.32	5.4
omniASR-LLM-3B	LLM	0.10	0.35	8.9
omniASR-LLM-1B	LLM	0.10	0.37	9.4
omniASR-CTC-7B	Enc-only	0.11	0.39	128.0
omniASR-LLM-300M	LLM	0.12	0.40	15.9
omniASR-CTC-3B	Enc-only	0.12	0.43	235.4
omniASR-CTC-1B	Enc-only	0.14	0.48	383.7
omniASR-CTC-300M	Enc-only	0.16	0.54	639.3
mms-1b-all	Enc-only	0.16	0.52	303.6
mms-1b-fl102	Enc-only	0.20	0.52	306.9
paza-Phi-4-MM	LLM	0.46	0.96	40.4
parakeet-tdt-0.6b-v3	E-D	0.48	0.97	2181.6
wav2vec2-conformer	Enc-only	0.49	1.11	239.8
data2vec-audio-large	Enc-only	0.50	1.04	727.3

Table 2: Top 15 models ranked by median CER across all languages. Arch.: LLM = LLM-Based, Enc-only = Encoder-only, E-D = Encoder-Decoder.

are even more pronounced ($H = 1985.5$), with all pairwise Cliff’s $\delta > 0.75$, confirming that efficiency is architecture-driven rather than an artefact of model selection.

4.4 Model Performance by Language Family

Table 3 summarizes the distribution of languages across families along with median error rates. Figure 6 shows CER distributions aggregated by language family. Afrikaans, the sole Indo-European language in the benchmark, achieves the lowest error rates, while Afroasiatic and Nilo-Saharan language families exhibit higher median CER and greater vari-

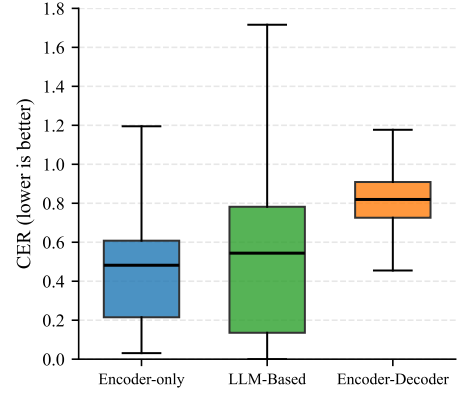


Figure 4: CER distribution by architectural paradigm.

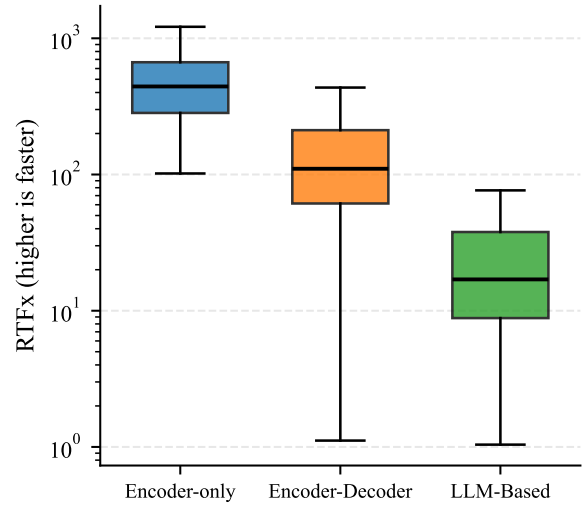


Figure 5: RTFx comparison by architecture (log scale).

ability, likely due to non-Latin scripts and known morphological complexity.

Language Family	Langs	CER _{med}	WER _{med}
Indo-European	1	0.561	0.963
Niger-Congo	26	0.665	0.989
Nilo-Saharan	3	0.790	0.998
Afroasiatic	9	0.964	1.000

Table 3: Language families with aggregate accuracy metrics.

4.5 Language Support

We group languages with lower median CER as most supported and those with higher CER as least supported. Figures 7 and 8 show the most and least supported languages based on median CER. Niger-Congo languages dominate the top 10 supported languages, with Afrikaans (Indo-European) and Oromo (Afroasiatic) as exceptions. Afroasiatic languages dominate the least supported languages, highlighting

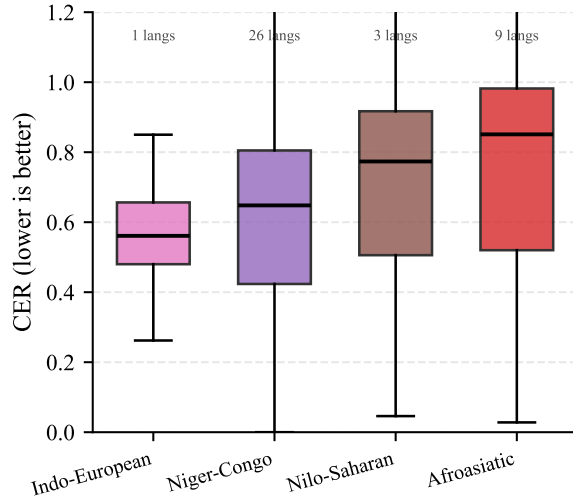


Figure 6: CER distribution across language families.

the need for script-aware ASR architectures and expanded training data for underrepresented orthographies.

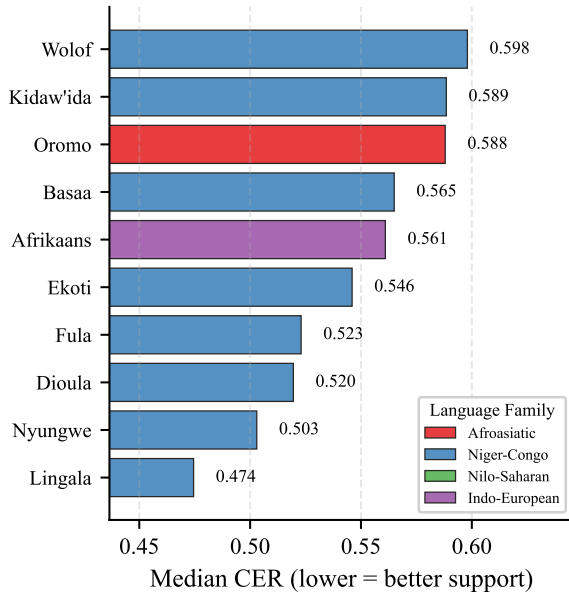


Figure 7: Top 10 most supported languages ranked by median CER.

4.6 Accuracy–Efficiency Trade-offs

Figure 9 illustrates the relationship between transcription accuracy and inference efficiency across all evaluated models. Encoder-only models occupy the high-throughput region, while LLM-based models cluster in the low-error, low-throughput region. Our analysis shows that inference efficiency is primarily architecture-driven rather than by model size as shown by Figure 11. Language dependence varies by decoding strategy (e.g., CTC models are more speed-invariant than autoregressive models)

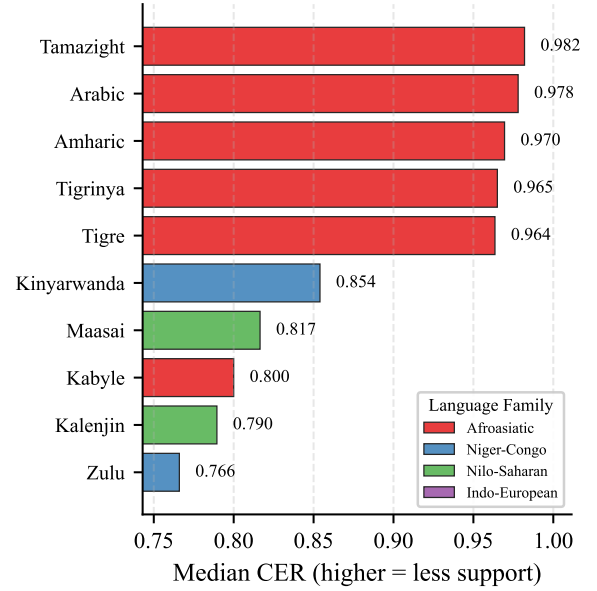


Figure 8: Top 10 least supported languages ranked by median CER.

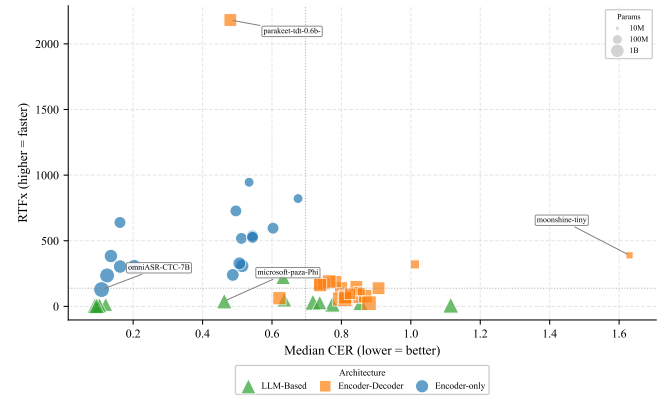


Figure 9: Accuracy–efficiency trade-off (CER vs. RTFx). Marker color indicates architecture; marker size reflects parameter count.

4.7 Relationship Between WER and CER

Figure 10 shows that WER and CER exhibit a strong positive correlation (Spearman $\rho = 0.64$, 95% bootstrap CI [0.61, 0.67]). Points above the line indicate disproportionately high character-level errors. A Kruskal-Wallis test on CER residuals from a linear fit ($R^2 = 0.67$) confirms significant family-level divergence ($H = 305.3$). Afroasiatic languages show the highest mean positive residual (+0.14), confirming that orthographic complexity drives disproportionately high character-level error rates, and that reporting both WER and CER is essential for multilingual ASR evaluation.

4.8 English-Only vs. Multilingual Models

Of the 52 evaluated models as seen on Table 8, 19 are trained exclusively on English data. To assess whether their inclusion biases conclusions, we repeat all statistical tests after excluding them (removing 1,154 of 3,168 evaluations, 36.5%). All

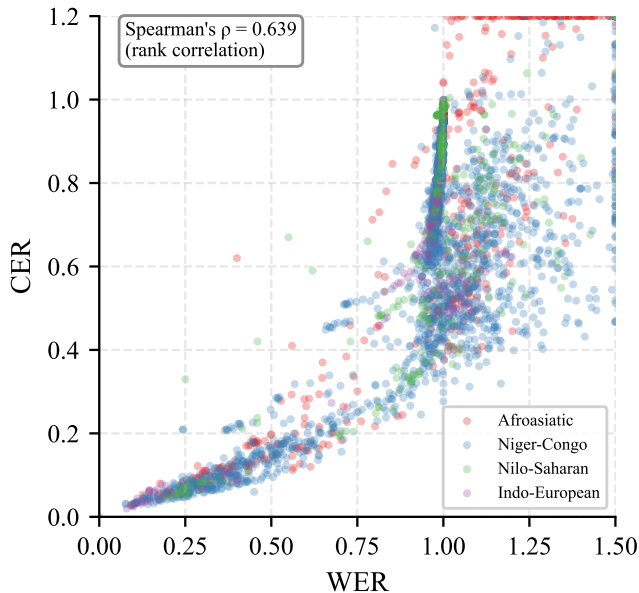


Figure 10: WER vs. CER correlation across all model-language pairs. Color indicates language family

conclusions remain significant ($p < 0.0001$) with effect sizes in the same interpretive range. Three effects become *stronger* after removal (architectural CER, WER-CER correlation, model family CER), while two weaken slightly (language family CER, language support gap). Notably, the Encoder-only vs. LLM-Based parity dissolves: Cliff’s δ shifts from -0.09 (negligible) to -0.40 (medium), revealing that apparent parity was partly driven by English-only models achieving moderate CER through English-word overlap rather than genuine multilingual capability.

5 Discussion

Architectural trade-offs. Statistically significant trade-offs emerge between accuracy and efficiency (Kruskal-Wallis $p < 10^{-200}$, large effect sizes). Encoder-only models offer substantially higher inference efficiency through non-autoregressive decoding while remaining competitive in accuracy where language support is consistent. LLM-based models achieve the lowest median error rates but incur higher computational cost. These differences highlight that architecture, not just parameter size, determines real-world ASR performance.

Accuracy-efficiency trade-off. Models achieving the lowest error rates generally incur higher computational cost. NVIDIA’s parakeet-tdt-0.6b-v3, despite being trained on English-only data, illustrates that architectural optimization can achieve high efficiency without proportional accuracy loss.

Language family differences. Afroasiatic and Nilo-Saharan languages exhibit higher error rates and greater variability than Niger-Congo languages, consistent with differences in morphology and non-Latin scripts. This suggests that future multilingual ASR systems would benefit from

stronger script-aware modelling and tokenization choices, rather than assuming a uniform multilingual decoding setup.

WER vs CER metrics. Although strongly correlated ($\rho = 0.64$), residual analysis confirms that Afroasiatic languages exhibit systematically higher CER than predicted by their WER ($p < 10^{-65}$). The substantial gap between WER and CER observed for certain languages indicates that character-level evaluation remains more informative for languages with richer morphology or non-Latin script.

English-only model impact. All benchmark conclusions hold after excluding 19 English-only models. The WER-CER correlation strengthens substantially ($\rho: 0.64 \rightarrow 0.81$), and the Encoder-only vs. LLM-Based parity dissolves ($\delta: -0.09 \rightarrow -0.40$), suggesting that multilingual coverage, not just architecture, determines cross-lingual accuracy.

6 Limitations

Data leakage. Public datasets used for evaluation may overlap with pre-training data used by large-scale ASR models introducing potential data leakage. Since model providers do not always disclose full training data composition, benchmark results should be interpreted as relative performance comparisons under shared evaluation conditions rather than absolute generalization estimates.

Limited language coverage. The benchmark does not capture Africa’s full linguistic diversity. Languages with extremely limited public data remain underrepresented. Future work will prioritize expanding coverage.

Potential language-model contamination. LLM-based ASR models may benefit from textual pretraining overlap which may provide benefit to why LLM-based models display higher accuracy.

Additional analysis. The benchmark focuses on transcription accuracy and inference efficiency, and does not explicitly evaluate robustness to domain shift, noise, or demographic factors.

7 Conclusion and Future Work

This benchmark’s findings underscore the significance of a large-scale low and mid-resource language ASR benchmark that integrates architectural analysis, deployment-relevant trade-offs on accuracy and efficiency metrics, and linguistically grounded evaluation. Together, they provide practical foundations for inclusive ASR development for diverse low-resource languages.

Future Work. Future work will broaden language coverage to other low-resource regions to enable cross-regional comparison, incorporate speaker and language identification, evaluate controlled fine-tuning and model adaptations, and assess real-world model robustness to noise, code-mixing/switching and domain shifts through data augmentation. We also plan to release inference outputs with detailed insertion, deletion, and substitution error analyses.

By providing a transparent and extensible benchmark, this work lays the groundwork for more principled, deployment-aware multilingual ASR research and continued progress toward inclusive speech technologies.

Appendix

A Evaluation Datasets

Table 4 summarizes the six publicly available datasets unified to construct the evaluation benchmark, capturing 61 dataset–language test splits across 39 African languages, totaling 199,438 evaluation samples.

Dataset	Description	Langs	Samples
Mozilla Common Voice 23.0	Crowdsourced speech	23	78,546
Naija Voices	Conversational speech	3	57,529
African Next Voices Kenya	Conversational speech	5	24,508
African Next Voices South Africa	Conversational speech	6	21,692
Google FLEURS	Read speech	20	12,813
ALFFA	Broadcast/read speech	3	4,350
Total		39	199,438

Table 4: Summary of the six evaluation datasets spanning 39 African languages.

B Model Size vs Inference Efficiency

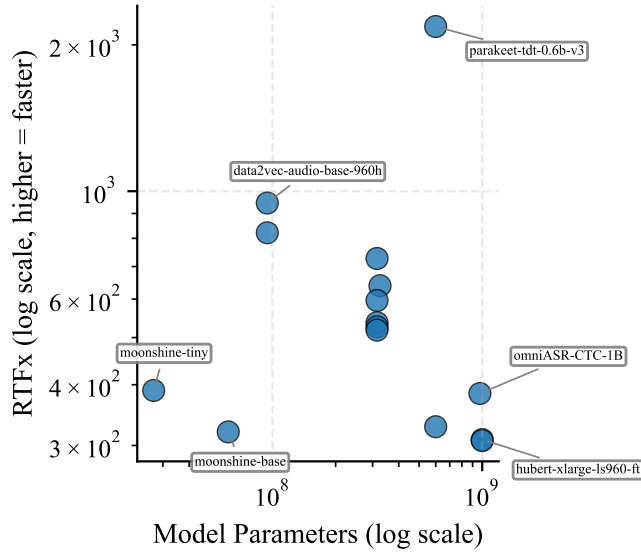


Figure 11: Model size vs. inference efficiency (RTFx) for the top 15 fastest models.

C Languages by Family

Table 5 shows the distribution of languages by family. Niger-Congo is the largest family with 26 languages (67%), pri-

marily from the Bantu branch. Afroasiatic contributes 9 languages (23%) spanning Semitic, Cushitic, Chadic, and Berber branches. Nilo-Saharan includes 3 Nilotic languages (8%), and Indo-European is represented by Afrikaans (3%).

Family	Language	Sub-family
<i>Afroasiatic (9 languages)</i>		
	Amharic	Semitic (Ethiopian)
	Arabic	Semitic (Central)
	Hausa	Chadic (West)
	Kabyle	Berber (Northern)
	Oromo	Cushitic (Lowland East)
	Somali	Cushitic (Lowland East)
	Tamazight	Berber (Northern)
	Tigre	Semitic (Ethiopian)
	Tigrinya	Semitic (Ethiopian)
<i>Indo-European (1 language)</i>		
	Afrikaans	Germanic (West)
<i>Niger-Congo (26 languages)</i>		
	Basaa	Bantu (A.40)
	Dioula	Mande (Western)
	Ekoti	Bantu (P.30)
	Fula	Atlantic (Senegambian)
	Igbo	Volta-Niger (Igboid)
	Kamba	Bantu (E.70)
	Kidaw'ida	Bantu (E.70 Taita)
	Kikuyu	Bantu (E.70)
	Kinyarwanda	Bantu (J.60)
	Lingala	Bantu (C.30)
	Luganda	Bantu (J.10)
	Northern Sotho	Bantu (S.30)
	Nyanja	Bantu (N.30)
	Nyungwe	Bantu (P.30)
	Sesotho	Bantu (S.30)
	Setswana	Bantu (S.30)
	Shona	Bantu (S.10)
	Swahili	Bantu (G.40)
	Tshivenda	Bantu (S.20)
	Twi	Volta-Niger (Kwa)
	Umbundu	Bantu (R.10)
	Wolof	Atlantic (Senegambian)
	Xhosa	Bantu (S.40)
	Xitsonga	Bantu (S.50)
	Yoruba	Volta-Niger (Defoid)
	Zulu	Bantu (S.40)
<i>Nilo-Saharan (3 languages)</i>		
	Dholuo	Nilotic (Western)
	Kalenjin	Nilotic (Southern)
	Maasai	Nilotic (Eastern)

Table 5: Complete listing of 39 African languages organized by language family.

D Full Model Performance Summary

Table 8 lists all 52 evaluated models ranked by median CER across languages, with multilingual and English-only models separated. Notably, 33 models are multilingual while 19 are trained exclusively on English data (LibriSpeech 960h checkpoints, Whisper .en variants, Moonshine, Parakeet-TDT, and Kyutai-STT-en). The impact of this composition on benchmark conclusions is analyzed in Section 4

E Architecture and Model Family Summaries

Table 6 confirms systematic differences across the three major architectural paradigms. Table 7 ranks the top 15 individual models by inference efficiency (RTFx), highlighting the accuracy–efficiency trade-off at the model level

Architecture	CER _{med}	WER _{med}	RTFx _{med}	N
Encoder-only	0.50	1.03	450.9	16
LLM-Based	0.63	1.04	17.4	13
Encoder-Decoder	0.81	0.99	115.8	23

Table 6: Architecture-level performance summary (ranked by median CER).

Model	Arch.	Size	CER	WER	RTFx
parakeet-tdt-0.6b-v3	E-D	600M	0.48	0.97	2181.6
data2vec-audio-base-960h	Enc	94M	0.53	1.09	945.7
wav2vec2-base-960h	Enc	94M	0.68	1.08	821.5
data2vec-audio-large-960h	Enc	315M	0.50	1.04	727.3
omniASR-CTC-300M	Enc	325M	0.16	0.54	639.3
wav2vec2-large-960h	Enc	315M	0.60	1.10	596.1
wav2vec2-large-960h-lv60	Enc	315M	0.54	1.01	536.5
wav2vec2-large-robust-ft	Enc	315M	0.54	1.09	526.1
hubert-large-ls960-ft	Enc	315M	0.51	1.03	518.0
moonshine-tiny	E-D	27M	1.63	2.23	389.3
omniASR-CTC-1B	Enc	975M	0.14	0.48	383.7
wav2vec2-conformer-rope	Enc	600M	0.51	1.08	327.9
moonshine-base	E-D	62M	1.01	1.20	319.9
hubert-xlarge-ls960-ft	Enc	1.0B	0.51	1.03	308.5
mms-1b-fl102	Enc	1.0B	0.20	0.52	306.9

Table 7: Top 15 models ranked by inference efficiency (RTFx, higher is faster). Best CER and WER among these models are **bolded**. Encoder-only models dominate, but omniASR-CTC-300M uniquely combines high efficiency with low error rates.

Model	Arch.	Size	CER	WER	RTFx
<i>Multilingual models (33)</i>					
omniASR-LLM-7B	LLM	7.8B	0.09	0.34	7.5
omniASR-LLM-7B_ZS	LLM	7.8B	0.10	0.32	5.4
omniASR-LLM-3B	LLM	4.4B	0.10	0.35	8.9
omniASR-LLM-1B	LLM	2.3B	0.10	0.37	9.4
omniASR-CTC-7B	Enc	6.5B	0.11	0.39	128.0
omniASR-LLM-300M	LLM	1.6B	0.12	0.40	15.9
omniASR-CTC-3B	Enc	3.1B	0.12	0.43	235.4
omniASR-CTC-1B	Enc	975M	0.14	0.48	383.7
omniASR-CTC-300M	Enc	325M	0.16	0.54	639.3
mms-1b-all	Enc	1.0B	0.16	0.52	303.6
mms-1b-fl102	Enc	1.0B	0.20	0.52	306.9
paza-Phi-4-MM	LLM	5.6B	0.46	0.96	40.4
paza-whisper-large-v3-turbo	E-D	809M	0.62	0.96	62.4
canary-qwen-2.5b	LLM	2.5B	0.63	1.12	222.8
granite-speech-3.3-2b	LLM	3.0B	0.64	1.09	51.9
granite-speech-3.3-8b	LLM	9.0B	0.72	1.12	32.7
whisper-large-v3-turbo	E-D	809M	0.74	0.97	162.1
distil-whisper-large-v3	E-D	756M	0.74	0.98	168.8
lite-whisper-large-v3-turbo-acc	E-D	600M	0.75	0.97	174.9
canary-1b-v2	E-D	1.0B	0.77	1.29	189.4
Qwen2-Audio-7B-Instruct	LLM	8.0B	0.77	1.04	17.4
whisper-large	E-D	1.6B	0.80	0.98	56.0
distil-whisper-large-v2	E-D	756M	0.80	0.99	138.1
whisper-large-v3	E-D	1.6B	0.81	0.98	53.7
lite-whisper-large-v3-acc	E-D	1.0B	0.81	0.99	71.6
whisper-large-v2	E-D	1.6B	0.81	0.98	56.3
lite-whisper-large-v3-turbo	E-D	600M	0.84	0.99	149.4
Phi-4-MM	LLM	5.6B	0.85	1.23	27.2
lite-whisper-large-v3	E-D	1.0B	0.85	0.99	73.0
lite-whisper-large-v3-fast	E-D	1.0B	0.87	0.99	78.0
seamless-m4t-v2-large	E-D	2.3B	0.88	1.24	24.4
lite-whisper-large-v3-turbo-fast	E-D	600M	0.91	1.00	138.4
Qwen2-Audio-7B	LLM	8.0B	1.11	1.30	10.0
<i>English-only models (19)</i>					
parakeet-tdt-0.6b-v3	E-D	600M	0.48	0.97	2181.6
wav2vec2-conformer-rel-pos	Enc	600M	0.49	1.11	239.8
data2vec-audio-large-960h	Enc	315M	0.50	1.04	727.3
wav2vec2-conformer-rope-large	Enc	600M	0.51	1.08	327.9
hubert-large-ls960-ft	Enc	315M	0.51	1.03	518.0
hubert-xlarge-ls960-ft	Enc	1.0B	0.51	1.03	308.5
data2vec-audio-base-960h	Enc	94M	0.53	1.09	945.7
wav2vec2-large-960h-lv60	Enc	315M	0.54	1.01	536.5
wav2vec2-large-robust-ft	Enc	315M	0.54	1.09	526.1
wav2vec2-large-960h	Enc	315M	0.60	1.10	596.1
wav2vec2-base-960h	Enc	94M	0.68	1.08	821.5
kyutai-stt-2.6b-en	LLM	2.6B	0.74	1.19	29.7
distil-whisper-medium.en	E-D	394M	0.78	0.99	188.9
whisper-medium.en	E-D	769M	0.81	0.99	67.8
whisper-small.en	E-D	244M	0.82	0.99	95.7
whisper-base.en	E-D	73M	0.85	0.99	115.8
whisper-tiny.en	E-D	38M	0.86	1.00	111.3
moonshine-base	E-D	62M	1.01	1.20	319.9
moonshine-tiny	E-D	27M	1.63	2.23	389.3

Table 8: All 52 evaluated models ranked by median CER within each group. Best CER, WER, and RTFx are **bolded**. Arch.: LLM = Multitask LLM, Enc = Encoder-only, E-D = Encoder-Decoder.

References

- [Abouelenin *et al.*, 2025] Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, Xiren Zhang, and Xiren Zhou. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.
- [Alabi and others, 2025] Jesujoba O. Alabi *et al.* AfriHuBERT: A self-supervised speech representation model for African languages. In *Proc. Interspeech*, 2025.
- [Ardila *et al.*, 2020] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 4218–4222. European Language Resources Association (ELRA), 2020.
- [Babu *et al.*, 2022] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saber, Juan Carbonell, *et al.* XLS-R: Self-supervised cross-lingual speech representation learning at scale. In *Proc. Interspeech*, 2022.
- [Baevski *et al.*, 2020] Alexei Baevski, Yuhao 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*, volume 33, pages 12449–12460, 2020.
- [Baevski *et al.*, 2022] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In *Proceedings of the International Conference on Machine Learning*, volume 162, pages 1298–1312. PMLR, 2022.
- [Barrault *et al.*, 2023] Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenhaler, Paul-Ambroise Duquenne, Brian Ellis, Hady El-sahar, Justin Haaheim, *et al.* SeamlessM4T: Massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*, 2023.
- [Black, 2019] Alan W. Black. CMU wilderness multilingual speech dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.
- [Boito and others, 2024] Marcelly Zanon Boito *et al.* mHuBERT-147: A compact multilingual HuBERT model. In *Proc. Interspeech*, 2024.
- [Chen and others, 2024] William Chen *et al.* Towards robust speech representation learning for thousands of languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024.
- [Chorowski *et al.*, 2015] Jan K. Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. In *Advances in Neural Information Processing Systems*, 2015.
- [Chu *et al.*, 2024] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.
- [Cliff, 1993] Norman Cliff. Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, 114(3):494–509, 1993.
- [Conneau *et al.*, 2021] Alexis Conneau, Alexei Baevski, Roman Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition. In *Proc. Interspeech*, 2021.
- [Conneau *et al.*, 2022] Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. FLEURS: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2022.
- [Demšar, 2006] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.
- [Dror *et al.*, 2018] Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker’s guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 1383–1392, 2018.
- [Elmadany *et al.*, 2025] AbdelRahim A. Elmadany, Sang Yun Kwon, Hawau Olamide Toyin, Alcides Alcoba Inciarte, Hanan Aldarmaki, and Muhammad Abdul-Mageed. Voice of a continent: Mapping Africa’s speech technology frontier. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11028–11050, 2025.
- [Emezue *et al.*, 2025] Chris Emezue, NaijaVoices Community, Busayo Awobade, Abraham Owodunni, Handel Emezue, Gloria Monica Tobechukwu Emezue, Neferiti Nneoma Emezue, Sewade Ogun, Bunmi Akinremi, David Ifeoluwa Adelani, *et al.* The NaijaVoices

- dataset: Cultivating large-scale, high-quality, culturally-rich speech data for African languages. In *Proc. Interspeech*, 2025.
- [Gandhi *et al.*, 2023] Sanchit Gandhi, Patrick von Platen, and Alexander M. Rush. Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling. *arXiv preprint arXiv:2311.00430*, 2023.
- [Gauthier *et al.*, 2016] Elodie Gauthier, Laurent Besacier, Sylvie Voisin, Michael Melese, and Uriel Pascal Curiel. Collecting resources in sub-saharan african languages for automatic speech recognition. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, 2016.
- [Hsu *et al.*, 2021] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.
- [Javed and others, 2023] Tahir Javed et al. IndicSUPERB: A speech processing universal performance benchmark for Indian languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2023.
- [Jeffries *et al.*, 2024] Nat Jeffries, Evan King, Manjunath Kudlur, Guy Nicholson, James Wang, and Pete Warden. Moonshine: Speech recognition for live transcription and voice commands. *arXiv preprint arXiv:2410.15608*, 2024.
- [Kamahori *et al.*, 2025] Keisuke Kamahori, Jungo Kasai, Noriyuki Kojima, and Baris Kasikci. LiteASR: Efficient automatic speech recognition with low-rank approximation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3430–3442. Association for Computational Linguistics, 2025.
- [Lewis, 2009] M. Paul Lewis, editor. *Ethnologue: Languages of the World*. SIL International, Dallas, TX, USA, sixteenth edition, 2009.
- [Marivate *et al.*, 2025] Vukosi Marivate, Kayode Olaleye, Sitwala Mundia, Andinda Bakainga, Unarine Netshifhefhe, Mahmooda Milanzie, Tsholofelo Hope Mogale, Thapelo Sindane, Zainab Abdulrasaq, Kesego Mokgosi, Chijioke Okorie, Nia Zion Van Wyk, Graham Morrissey, Dale Dunbar, Francois Smit, Tsosheletso Chidi, Rooweither Mabuya, Andiswa Bukula, Respect Mlambo, Tebogo Macucwa, Idris Abdulmumin, , and Seani Rananga. Swivuriso: The south african next voices multilingual speech dataset. *arXiv preprint arXiv: 2512.02201*, 2025.
- [Omnilingual ASR Team *et al.*, 2025] Omnilingual ASR Team, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, Chierh Cheng, Joe Chuang, Caley Droof, Mark Duppenthaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, Kehan Lyu, Sagar Miglani, Vineel Pratap, Kaushik Ram Sadagopan, Safiyyah Saleem, Arina Turkatenko, Albert Ventayol-Boada, Zheng-Xin Yong, Yu-An Chung, Jean Maillard, Rashel Moritz, Alexandre Mourachko, Mary Williamson, and Shireen Yates. Omnilingual ASR: Open-source multilingual speech recognition for 1600+ languages. *arXiv preprint arXiv:2511.09690*, 2025.
- [Pratap *et al.*, 2020] Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. MLS: A large-scale multilingual dataset for speech research. In *Proc. Interspeech*, 2020.
- [Pratap *et al.*, 2024] 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. *Journal of Machine Learning Research*, 25(97):1–52, 2024.
- [Radford *et al.*, 2023] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the International Conference on Machine Learning*. PMLR, 2023.
- [Saon *et al.*, 2025] George Saon, Avihu Dekel, Alexander Brooks, Tohru Nagano, Abraham Daniels, Aharon Satt, Ashish Mittal, Brian Kingsbury, David Haws, Edmilson Morais, Gakuto Kurata, Hagai Aronowitz, Ibrahim Ibrahim, Jeff Kuo, Kate Soule, Luis Lastras, Masayuki Suzuki, Ron Hoory, Samuel Thomas, Sashi Novitasari, Takashi Fukuda, Vishal Sunder, Xiaodong Cui, and Zvi Kons. Granite-speech: Open-source speech-aware llms with strong english asr capabilities. *arXiv preprint arXiv:2505.08699*, 2025.
- [Shi *et al.*, 2023] Jiatong Shi, Dan Berrebbi, William Chen, En-Pei Hu, Wei-Ping Huang, Ho-Lam Chung, et al. ML-SUPERB: Multilingual speech universal PERFORMANCE benchmark. In *Proc. Interspeech*, pages 884–888, 2023.
- [Srivastav *et al.*, 2025] Vaibhav Srivastav, Steven Zheng, Eric Bezzam, Eustache Le Bihan, Nithin Koluguri, Piotr Żelasko, Somshubra Majumdar, Adel Moumen, and Sanchit Gandhi. Open asr leaderboard: Towards reproducible and transparent multilingual and long-form speech recognition evaluation, 2025.
- [Wang *et al.*, 2025] Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F Chen. Audiobench: A universal benchmark for audio large language models. *NAACL*, 2025.
- [Wanzare *et al.*, 2026] Lilian Wanzare, Cynthia Amol, Ezekiel Maina, Nelson Odhiambo, Hope Kerubo, Leila Misula, Vivian Oloo, Rennish Mboya, Edwin Onkoba, Edward Ombui, Joseph Muguro, Ciira wa Maina, Andrew Kipkebut, Alfred Omondi Otom, Ian Ndung’u Kang’ethe, Angela Wambui Kanyi, and Brian Gichana Omwenga. Afrivoices-ke: A multilingual speech dataset for kenyan languages, 2026.
- [Xu *et al.*, 2023] Hainan Xu, Fei Jia, Somshubra Majumdar, He Huang, Shinji Watanabe, and Boris Ginsburg. Efficient

sequence transduction by jointly predicting tokens and durations. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 38462–38484. PMLR, 2023.

[Zeghidour *et al.*, 2025] Neil Zeghidour, Eugene Kharitonov, Manu Orsini, Václav Volhejn, Gabriel de Marmiesse, Edouard Grave, Patrick Pérez, Laurent Mazaré, and Alexandre Défossez. Streaming sequence-to-sequence learning with delayed streams modeling. Technical report, 2025.