AfriSpeech-MultiBench: A Verticalized Multidomain Multicountry Benchmark Suite for African Accented English ASR

Gabrial Zencha Ashungafac[1], Mardhiyah Sanni[1], Busayo Awobade[1], Alex Gichamba[1], Tobi Olatunji[1]

[1] Intron Health

tobi@intron.io

Abstract

Recent advances in speech-enabled AI, including Google's NotebookLM and OpenAI's speech-to-speech API, are driving widespread interest in voice interfaces globally. Despite this momentum, there exists no publicly available application-specific model evaluation that caters to Africa's linguistic diversity. We present **AfriSpeech-MultiBench**, the first domain-specific evaluation suite for over 100 African English accents across 10+ countries and seven application domains: Finance, Legal, Medical, General dialogue, Call Center, Named Entities and Hallucination Robustness. We benchmark a diverse range of open, closed, unimodal ASR and multimodal LLM-based speech recognition systems using both spontaneous and non-spontaneous speech conversation drawn from various open African accented English speech datasets. Our empirical analysis reveals systematic variation: open-source ASR models excels in spontaneous speech contexts but degrades on noisy, non-native dialogue; multimodal LLMs are more accent-robust yet struggle with domain-specific named entities; proprietary models deliver high accuracy on clean speech but vary significantly by country and domain. Models fine-tuned on African English achieve competitive accuracy with lower latency, a practical advantage for deployment, hallucinations still remain a big problem for most SOTA models. By releasing this comprehensive benchmark, we empower practitioners and researchers to select voice technologies suited to African use-cases, fostering inclusive voice applications for underserved communities.

1 Introduction

Automatic Speech Recognition (ASR) has become a foundational technology across numerous domains. In customer-support environments, ASR powers real-time call routing, intent detection, and agent assistance, substantially reducing response times and improving user satisfaction (Wang et al., 2023). In healthcare, voice-enabled digital scribes transcribe clinician-patient interactions on the fly, alleviating documentation burdens and cutting downstream transcription costs (van Buchem et al., 2021). Emerging applications in legal transcription (Saadany et al., 2023), financial trading desktops, and live subtitling further demonstrate the broad impact of ASR systems in both enterprise and consumer settings.

Selecting the optimal ASR model for a given task now often means choosing among powerful, pre-trained *foundation* systems rather than training bespoke models from scratch. Self-supervised models such as wav2vec 2.0 (Baevski et al., 2020) learn rich audio features from large amounts of unlabeled speech and can be applied in a zero-shot or few-shot manner, achieving near state-of-the-art WER rates on standard benchmarks (Baevski et al., 2020). Large multitask models such as Whisper (Radford et al., 2023), trained on hundreds of thousands of hours of multilingual and multitask data, exhibit strong zero-shot transfer across domains and languages without additional fine-tuning (Radford et al., 2023). However, computational budgets, latency requirements, and domain mismatches mean that one foundation model may outperform another depending on the target task, be it medical dictation, legal proceedings, or informal conversational speech.

Accented speech, particularly non-Western and underrepresented varieties, remains a persistent blind spot in mainstream evaluation suites. African accents, with their rich phonetic and prosodic diversity, often lead to significant word error rate disparities compared to North-American or British English (Dossou, 2025). Without a dedicated benchmark, practitioners lack a reliable way to assess which off-the-shelf ASR system can meet accuracy, latency, or robustness requirements on African-accented speech.

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 3642–3653

December 20-24, 2025 ©2025 Association for Computational Linguistics

Accordingly, we present a unified evaluation suite that benchmarks leading ASR systems, **AfriSpeech-MultiBench** in zero-shot mode across medical, legal, conversational, entity-rich, call center, finance and robustness-diagnostic African-accented English speech (short, silence, and no-speech conditions). The suite supplies standardized test sets, and transparent scoring protocols enabling practitioners to compare models and select the architecture most appropriate for their target application or for finetuning. We publicly release the benchmark suite on Hugging Face with CC BY-NC-SA 4.0 License [1]

2 Related Work

IrokoBench introduced a comprehensive text-based evaluation across seventeen low-resource African languages, revealing significant performance gaps between large language models and human competence on tasks such as natural language inference, reasoning and question answering (Adelani et al., 2025). The study underscores the necessity of domain-specific evaluation: without targeted test suites, systematic deficiencies remain undetected.

Within automatic speech recognition (ASR), progress is often measured through the community-maintained Open ASR Leaderboard, which continuously reports word-error rate (WER) and real-time factor on LibriSpeech (Panayotov et al., 2015), TED-LIUM 3 (Hernandez et al., 2018), GigaSpeech (Chen et al., 2021), VoxPopuli (Wang et al., 2021), AMI (Carletta et al., 2005), Earnings22 (Andrew et al., 2022), SPGISpeech (Guo et al., 2022), and Common Voice (Ardila et al., 2020). Although these datasets cover a range of domains, from read audiobooks to meeting-room recordings, they remain dominated by North-American and British English, providing limited insight into performance on African-accented English.

Empirical investigations confirm the practical consequences of this imbalance. Koenecke et al., 2020 documented a twofold increase in WER for African American Vernacular English relative to Standard American English across multiple commercial recognizers. A global audit involving speakers from 171 birth countries observed the largest error rates for sub-Saharan participants(DiChristofano et al., 2022). In the absence of African-accented evaluation sets, leaderboard rankings therefore offer an incomplete picture for stakeholders on the continent.

Modern recognizers are architecturally diverse. They include multilingual encoders such as Whisper (Radford et al., 2023) and XLS-R, proprietary cloud services (Microsoft Azure Speech-to-Text, Google Speech-to-Text), Conformer-based systems like Canary (Puvvada et al., 2024) and Parakeet (Rekesh et al., 2023), Speech-Augmented Language Models (SALMs) (Chen et al., 2023), and multimodal architectures such as SeamlessM4T (Schwenk et al., 2023). Their heterogeneous training regimes and objectives complicate any attempt to infer accent robustness from results on existing benchmarks alone.

Several African-accented corpora have been released to mitigate data scarcity. AfriSpeech-200 provides roughly 200 hours of read speech from more than 100 indigenous accents (Olatunji et al., 2023). AfriSpeech-Dialog adds spontaneous two-speaker conversations (Sanni et al., 2025); AfriSpeech-Parliament captures parliamentary debates (Intron Health, 2025a); Med-Convo-Nig focuses on Nigerian clinical tele-consultations (Intron Health, 2025c); Afri-Names targets named-entity-rich prompts (Intron Health, 2025b); and AfriSpeech-Countries assembles cross-regional accents under consistent recording conditions (Intron Health, 2025). Existing baseline evaluations do not cover modern speech recognition systems or lack broad application-specific results.

This study contributes three key advances. First, 7 publicly available African-accented corpora are harmonised into AfriSpeech-MultiBench, an evaluation suite spanning medical, legal, conversational , named-entity-rich and noise robustness speech. Second, 18 contemporary recognizers covering multilingual, proprietary, Conformer-based, SpeechLLMs and multimodal architectures are evaluated in zero-shot mode, with WER reported. Third, a fine-grained error analysis disaggregates results by accent cluster, phonetic context and domain, elucidating systematic failure modes and informing future data collection and model selection.

[1] https://huggingface.co/collections/intronhealth/afrispeechmultibench

3 Benchmark Methodology

3.1 Source Datasets

We assemble seven corpora to form AfriSpeech-MultiBench, covering diverse Anglophone African English accents. The distribution of sources is shown in Table 1.

- **AfriSpeech-200**: (Afri) a 200-hour, 67,577 clip dataset, 2,463 speakers across 120 indigenous accents from 13 African countries, spanning clinical and general domain read speech (Olatunji et al., 2023).
- **AfriSpeech-Dialog:** (Diag) about 50 long-form medical and nonmedical conversational sessions with African-accented spontaneous English (about 7 hrs) (Sanni et al., 2025).
- **AfriSpeech-Parliamentary:** (Parl) A real-world noisy, multi-speaker dataset of transcribed parliamentary speech (about 35.86 hours, 8,068 clips) sampled from Nigeria, Ghana, South Africa, and Kenya. (Intron Health, 2025a).
- **Med-Conv-Nig:** (Med.Conv) about 25 long-form simulated doctor-patient conversations capturing multispecialty clinical interactions in Nigeria, featuring both male and female speakers and rich in medical vocabulary tailored for evaluating domain-specific ASR in healthcare settings (Intron Health, 2025c).
- **AfriNames:** (Names) A read-speech corpus with subsets focused on African names (Name), numbers (Nums), and voice commands (Commands), e.g. "transfer $500 to my HSBC account"; comprising 6,307 single-speaker samples (about 8.92 hours), enriched with named entities and number utterances, spanning 12 distinct accents across four countries, particularly suited for evaluating ASR performance on entity-rich transcription tasks (Intron Health, 2025b)
- **AfriSpeech-Countries:** A mixture of AfriSpeech-200, AfriSpeech-Parliamentary, AfriNames and North African accented speech samples (Ctry-NA), totaling approximately 67 hours and 21,581 clips. The dataset spans seven African regions and includes both read and conversational speech. All samples are annotated by domain and country.
- **Afro-Call-Centers:** (Call) A private unreleased dataset capturing real-world agent-customer voice interactions rich in domain-specific vocabulary across finance, health, and customer support domains

3.2 Domains Studied

We define seven domain categories for evaluation with dataset details described in Table 1:

- **Medical:** health-related medical speech and clinician–patient dialogues.
- **General:** read-speech sourced from Wikipedia and non-spontaneous multispeaker dialogues.
- **Legal:** noisy parliamentary proceeding with overlapping speech.
- **Finance:** read speech enriched with numbers such as currencies, decimals, dates, measurements, locations, trading volumes, and financial institutions.
- **Call Center / Customer Support:** real-world agent–customer interactions
- **Named-Entities:** Named-Entity-Rich General clips with dense mentions of African person names, locations, organizations, and dates
- **Noise Robustness:** This diagnostic subset evaluates ASR stability under challenging acoustic conditions, including short utterances (under 3.5 s) from AMI, VoxPopuli, AfriSpeech, and AfriNames datasets. It also includes *Intervening Silence* clips from AfriNames with deliberate pauses to test contextual continuity, and a *No Speech* subset from AfriSpeech-Parliament to measure false-trigger resistance when no speech is present.

3.3 Models

We evaluate 18 modern ASR systems partly sourced from the top twenty entries on the Hugging Face Open ASR Leaderboard (snapshot: July 2025)[2] categorized into model families representing architectural breadth Conformer, RNN-T, CTC, transducer hybrids, and speech-augmented language models (SpeechLLMs) and include both fully open-source checkpoints and proprietary services already deployed in commercial workflows.

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

Domain	Data Source	Samples	Hours	Countries	Accents	Speakers
Medical	Afri (clinical), Dialog (medical), Med.Conv	3651	29.88	10	95	519
General	Afri (general), Dialog (general)	2741	13.06	9	84	455
Legal	Parl	8068	35.86	4	–	–
Named Entities	Names (names)	3121	2.18	3	6	–
Finance	Names (numbers), Names (commands)	3186	6.73	4	9	–
Call Center	Call (Private)	16	0.80	2	3	32
Robustness	Short Speech, No Speech, Intervening Silence	2067	3.74	-	-	-
Total Unique		**20093**	**79.19**	**11**	**108**	**859**

Table 1: Domain-wise breakdown of the AfriSpeech-Multibench benchmark. Parentheses denote domain-specific subsets. Full names of the datasets - Afri:AfriSpeech, Dialog:AfriSpeech-Dialog, Med.Conv:Med-Conv-Nig, Names: AfriNames. The Call Center source is private and not disclosed.

Architecture	Model	Size
Conformer	Nvidia Parakeet-tdt-0.6B-v2	0.6B
	Nvidia Parakeet-tdt-1.1B	1.1B
	Nvidia Parakeet-rnnt-1.1B	1.1B
	Nvidia Canary-1B-flash	1B
Whisper	OpenAI Whisper-large-v3	1.54B
	Distil-Whisper-v3.5	756M
	Nyra Health CrisperWhisper	1.54B
SpeechLLMs	IBM Granite-3.3-2B	2B
	Mistral Voxtral-Mini-3B	3B
	Nvidia Canary-Qwen-2.5B	2.5B
	Microsoft Phi-4 MM-Instruct	5.6B
Proprietary	Intron-Sahara	–
	Intron-Sahara-V2	–
	OpenAI GPT-4o Transcribe	–
	Google Gemini-2.0 Flash	–
	AWS Transcribe	–
	Microsoft Azure Speech	–
	Google Chirp v3	–

Table 2: Descriptions of evaluated models, including model size, core architecture, and provider. Model sizes are in billions (B) of parameters when known.

- **NVIDIA's open models:** Open-source ASR models based on the FastConformer (Rekesh et al., 2023) such as the Parakeet variants: CTC, RNN-T and TDT (Galvez et al., 2024) in sizes of 0.6B and 1.1B, and the 1 billion parameter Canary-flash model pairing a FastConformer encoder with a transformer decoder (Puvvada et al., 2024).

- **Whisper Variants:** Transformer encoder decoder models based on Whisper (Radford et al., 2023). We consider the variants: Whisper-large-v3 (Radford et al., 2023), Distil-Whisper-v3.5[3], and CrisperWhisper (Zusag et al., 2024).

- **Open SpeechLLMs:** Multimodal LLMs and Speech-Augmented LLMs including IBM Granite-3.3-2B[4], Phi-4 Multimodal Instruct (Abdin et al., 2024), Nvidia Canary-Qwen[5], and Mistral's Voxtral Mini-3B (Liu et al., 2025).

- **Proprietary cloud ASR services:** OpenAI's GPT-4o transcribe[6], Google's Gemini-2.0-flash[7], Google's Chirp V3 [8], AWS Transcribe[9], Azure Speech Recognition[10] and Intron Sahara (V1 and V2)[11]. Models are evaluated in zero-shot mode, with neither demonstrations (Min et al., 2022) nor domain-specific fine-tuning.

This broad selection of modern ASR systems facilitate an empirical comparison between commercially deployed services and publicly available checkpoints, capturing the architectural and commercial diversity of leading ASR systems, providing a realistic basis for accent-aware model selection.

3.4 Evaluation Protocol

- Primary metric: Word Error Rate (WER) measured per model, per domain, per country, and per dataset.

- Error analysis: Breakdown by domain, accent group (native vs non-native), named-entity er-

[3] https://huggingface.co/distil-whisper/distil-large-v3.5

[4] https://huggingface.co/ibm-granite/granite-speech-3.3-2b

[5] https://huggingface.co/nvidia/canary-qwen-2.5b

[6] https://platform.openai.com/docs/models/gpt-4o-transcribe

[7] https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash

[8] https://cloud.google.com/speech-to-text/v2/docs/chirp_3-model

[9] https://aws.amazon.com/transcribe/

[10] https://azure.microsoft.com/en-us/products/ai-services/ai-speech

[11] https://www.intron.io/

rors, noise robustness; Open-source vs proprietary models, unimodal vs multimodal, large vs compact variants.

4 Experiments

- Dataset splits: We use held-out test sets per corpus, ensuring some accents appear only in testing to evaluate zero-shot generalization (e.g. 41 accents exclusively in test partition of AfriSpeech-200)
- Transcript Pre- and Post-processing: Model-specific transcript pre- and post-processing (described in Appendix section 7) normalized inputs, removed filler words, and mapped number words to their digit form, e.g. "twenty" to "20" and "first" to "1st".
- Inference setup: Uniform audio input preprocessing (16 kHz mono, no diarization) with default hyperparameters and decoding settings for ASR models and proprietary API calls. Local runs were on single T4 GPU (16GB).
- Prompting: We use consistent prompts for open and closed LLMs, e.g. "Transcribe this ENGLISH audio". Prompt details are provided in Appendix section 7.

We provide results for single runs.

5 Results

5.1 Overall Results

As shown in Table 3, model performance on widely used ASR benchmarks such as LibriSpeech, TED-LIUM, and AMI does not reliably translate to accuracy on African-accented or domain-specific speech. Models that dominate global leaderboards exhibit substantial degradation under accent, noise, and contextual variability characteristic of African datasets. For instance, leading open-source systems like **Parakeet-tdt-0.6B-v2** and **Whisper-large-v3**, which achieve sub-4% WER on LibriSpeech, experience error rates between 30-45% on general African speech and exceed 70% on medical dialogue or named-entity-rich inputs within **AfriSpeech-MultiBench**.

This discrepancy holds consistently across architecture families including transformer-based, transducer, and instruction-tuned multimodal models highlighting a persistent generalization gap of roughly 2-5× between leaderboard metrics and African deployment conditions. Notably, these errors compound in domain-specific contexts where pronunciation diversity, code-switching, and out-of-vocabulary proper nouns are prevalent.

In contrast, **Intron-Sahara V2** a successor to **Intron-Sahara** and a regionally tuned model trained on diverse African English data demonstrates markedly superior transferability. It achieves WERs below 15% across all benchmarked domains, with a modest increase to 18.09% in medical conversation, still outperforming all other models by a wide margin.

Beyond accuracy, **Intron-Sahara V2** exhibits exceptional robustness under challenging acoustic conditions. Across the robustness diagnostic subsets (*Silence*, *Short Samples*, and *Intervening Silence*), it consistently delivers the lowest error and false-trigger rates, underscoring its resilience to pause-filled, truncated, or low-energy speech. These findings suggest that regionally tuned ASR systems can close much of the performance gap on African speech, outperforming larger yet globally trained models across both accuracy and robustness dimensions.

5.2 Domain Performance

5.2.1 Medical

As shown in Table 4, the medical domain remains one of the most challenging settings, with average WERs exceeding 40% for most open and proprietary systems. **Intron-Sahara V2** achieves the best overall performance across all three medical datasets on *Afri-Med*, *Afri-Diag*, and *Med.Conv* yielding an overall average of 15.26%. Open models such as **Whisper-large-v3** and **Parakeet-tdt-0.6B-v2** perform moderately (25-27%), while proprietary systems like **Gemini-2.0**, **GPT-4o**, and **Azure** range from 24–30%. Large multimodal LLMs such as **Phi-4 MM-Instruct** and **IBM Granite** exceed 80% WER, underscoring that leaderboard success on clean English benchmarks does not generalize to accented or domain-specific medical speech. These results highlight the advantage of regionally tuned, domain-adapted ASR systems for low-resource healthcare applications in Africa.

5.2.2 Finance

The finance domain-represented by the *Afri-Names* subsets for numerals and spoken commands shows strong gains for regionally tuned ASR. **Intron-Sahara V2** achieves a WER of 8.20%, compared to 35-50% for most open-source and proprietary

Model	Open ASR Benchmarks							AfriSpeech-MultiBench						
	Lib-S	TED-3	Giga	VoxP	AMI	Earn22	SPGI	Afri	Diag	Parl	MedC	Names	Call	Rob
Parakeet-tdt-0.6B-v2	1.69	3.38	9.74	5.95	11.16	11.15	2.17	30.20	**11.23**	18.45	29.41	41.88	20.96	40.89
Parakeet-tdt-1.1B	**1.40**	3.59	9.52	5.49	15.87	14.49	3.16	28.45	15.14	27.14	29.98	45.66	25.26	44.62
Parakeet-rnnt-1.1B	1.45	3.83	9.89	**5.44**	17.01	13.94	2.93	28.18	15.08	26.75	30.59	46.70	28.93	90.30
Canary-1B-flash	1.48	3.12	9.85	5.63	13.11	12.77	1.95	29.77	48.50	19.13	93.62	44.10	88.71	51.32
Whisper-large-v3	2.01	3.86	10.02	9.54	15.95	11.29	2.94	26.49	13.49	19.99	31.76	43.23	24.69	33.79
Distil-Whisper-v3.5	2.37	3.64	9.84	8.04	14.63	11.29	2.87	27.58	11.50	18.00	30.41	45.80	21.65	34.00
CrisperWhisper	1.82	3.20	10.24	9.82	**8.71**	12.89	2.70	63.80	72.72	79.35	83.12	70.14	35.52	38.82
IBM Granite-3.3-2B	1.64	4.12	11.05	6.55	10.22	13.86	3.96	34.38	99.59	20.67	96.30	49.51	27.10	45.86
Voxtral (Mistral)	1.86	–	10.04	6.78	–	12.18	2.04	20.17	68.42	21.10	78.73	49.36	29.20	43.51
Canary-Qwen-2.5B	1.61	**1.90**	**9.43**	5.66	10.19	**10.45**	**1.90**	29.87	96.64	18.18	97.89	42.91	39.09	41.15
Phi-4 MM-Instruct	1.68	2.89	9.77	5.93	11.45	10.50	3.11	26.48	88.91	36.73	130.17	44.28	24.99	122.45
Intron-Sahara	–	–	–	–	–	–	–	16.35	14.26	15.41	27.92	**8.17**	20.08	18.91
Intron-Sahara V2	–	–	–	–	–	–	–	**11.83**	12.02	**13.01**	**18.09**	11.66	**13.45**	**7.86**
GPT-4o Transcribe	–	–	–	–	–	–	–	24.66	15.03	64.39	30.80	52.49	23.20	33.59
Google Gemini-2.0 Flash	–	–	–	–	–	–	–	27.80	12.02	20.51	27.59	50.12	22.39	33.91
AWS Transcribe	–	–	–	–	–	–	–	32.77	14.02	18.50	30.08	36.70	23.51	33.98
Azure Speech Recognition	–	–	–	–	–	–	–	28.41	13.29	18.75	29.17	35.69	24.95	32.61
Google Chirp V3	–	–	–	–	–	–	–	35.03	17.53	28.18	38.57	52.45	29.60	31.70

Table 3: Word Error Rate (WER %) for each model on standard open ASR benchmarks and subsets of the AfriSpeech-MultiBench dataset. Dashes represent results that were not available. Full names of datasets: Lib-S: LibriSpeech; TED-3: TED-LIUM 3; Giga: GigaSpeech; VoxP: VoxPopuli; AMI: AMI Meeting Corpus; Earn22: Earnings22; SPGI: SPGISpeech; Afri: AfriSpeech-200; Diag: AfriSpeech-Dialogue; Parl: AfriSpeech-Parliamentary; MedC: Med-Conv-Nig; Nam: AfriNames; Call: Afro-Call-Centers.; Robustness a combination of No Speech from Afrispeech-Parliamentary; Short Samples from AMI, VoxP and Afrispeech-Names; and Intervening Silence samples from AMI, VoxP, Afrispeech-Names

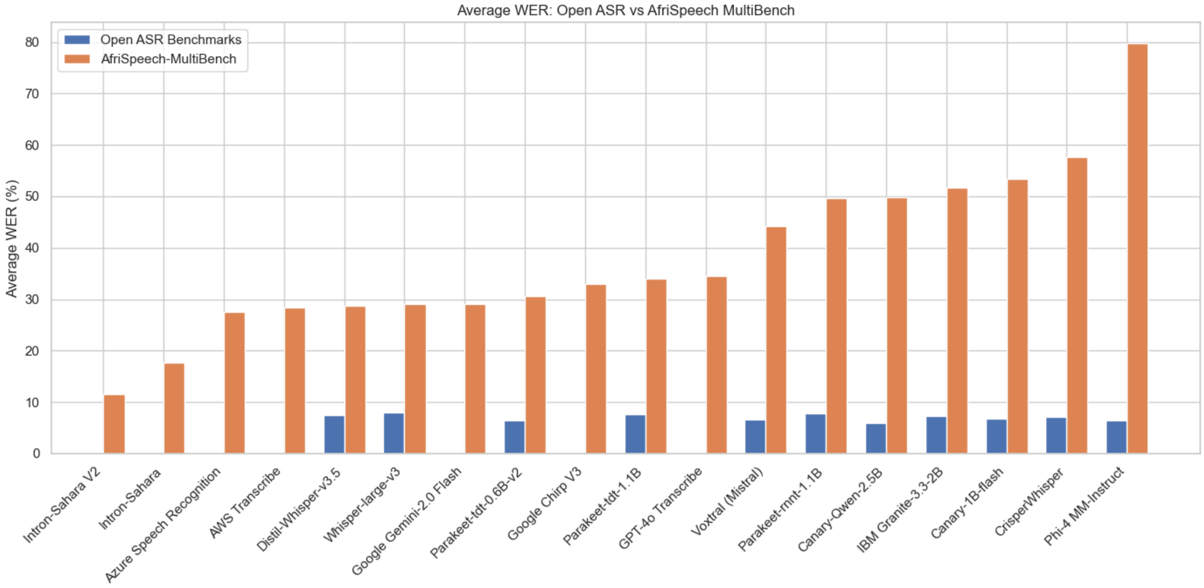

Figure 1: Average of Open ASR Leaderboard vs AfriSpeech-Multibench

systems. Lightweight conformer models such as **Parakeet-tdt-1.1B** and **Whisper-large-v3** reach around 43-46%, while **Azure** and **AWS** achieve mid-30% accuracy. This domain highlights Sahara's superior handling of short, context-free utterances and accent-driven variations in number pronunciation critical in the financial domain.

5.2.3 Names

Performance on African named entities remains a major differentiator. As shown in Table 3, **Intron-Sahara V2** again leads with 12.4% WER, compared to 40-70% for open-source models and over 50% for proprietary systems. The failure modes of large general-purpose LLMs (e.g., hallucinating or anglicizing names) emphasize the need for phonetic grounding and localized lexicons. Despite limited scale, Sahara's region-specific acoustic and language modeling yields more accurate entity recovery.

5.2.4 Legal

Table 3 summarizes performance on the *Parliamentary* dataset, which features overlapping speakers and high background noise. Among open-source systems, **Distil-Whisper-v3.5** performs best with

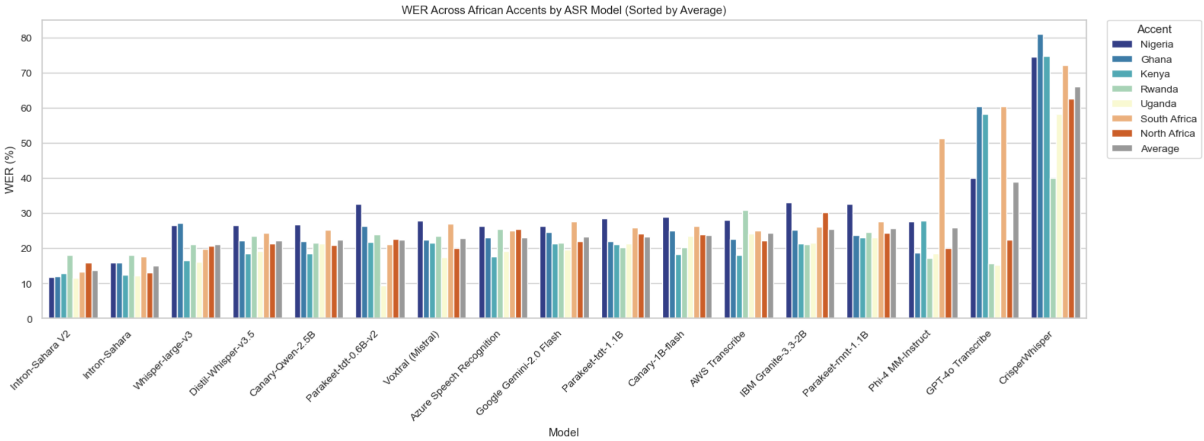

Figure 2: Word Error Rate (WER %) for each model across different African English accents in AfriSpeech-MultiBench. The average is computed across all listed accent categories.

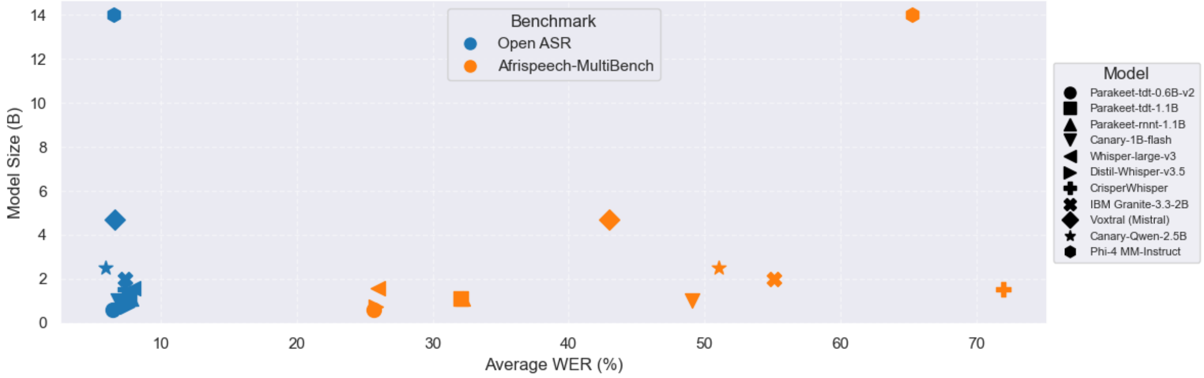

Figure 3: Model Sizes vs. Performance on Open ASR Benchmark (Blue) and AfriSpeech-Multibench (Orange)

11.5% WER, showing that smaller distilled variants can generalize better to real conversational overlap. **Intron-Sahara V2** follows closely at 12.01%, outperforming all proprietary and large multimodal models, whose WERs remain above 20%. This demonstrates that domain-adapted encoders trained on regional conversational data can surpass even the largest general-purpose models under high-noise conditions.

5.2.5 Call Center

In the call-center domain, **Intron-Sahara V2** again leads, achieving 13.45% WER, followed by **Whisper-large-v3** at 24.7% and **Parakeet-tdt-0.6B-v2** at 20.96%. Proprietary models such as **GPT-4o Transcribe** and **Gemini-2.0** record around 22-23% WER, showing that Sahara's tuning for multi-speaker interaction, turn-taking, and accent robustness provides a tangible advantage in noisy, dialogue-heavy environments.

5.2.6 Noise Robustness

The **Noise Robustness** evaluation combines the *Silence*, *Short*, and *Intervening Silence* diagnostics to assess model stability under pauses and non-speech conditions. Most global ASR systems degrade sharply (20–70% WER), often hallucinating speech. In contrast, **Intron-Sahara V2** achieves 0 % false triggers on silence, 10.15% on short clips, and 11.23% under pauses, outperforming *Whisper-large-v3* and *GPT-4o*. Detailed results are provided in Appendix 7

5.3 Accent and country variations

As shown in Table 6 and Figure 2, most models show pronounced degradation in Nigeria, South Africa, and Ghana (about 30%), relative to East and North Africa (about 24%). Most models perform comparably except GPT-4o and CrisperWhisper with WERs above 60% and Sahara models with WER less than 15%.

Model	**Afri-Med**	**Diag**	**Med.Conv**	**Average**
Parakeet-tdt-0.6B-v2	34.55	**11.23**	29.41	25.06
Parakeet-tdt-1.1B	33.79	15.14	29.98	29.98
Parakeet-rnnt-1.1B	33.45	15.08	30.59	30.59
Canary-1B-flash	34.77	72.23	78.92	78.92
Whisper-large-v3	32.59	17.22	31.76	27.19
Distil-Whisper-v3.5	32.18	16.77	30.63	26.53
CrisperWhisper	66.66	78.92	83.12	76.23
IBM Granite-3.3-2B	40.28	99.53	96.30	78.70
Voxtral (Mistral)	30.75	56.32	78.73	55.27
Canary-Qwen-2.5B	32.04	93.08	97.92	74.35
Phi-4 MM-Instruct	31.74	88.91	130.17	83.61
Intron-Sahara	19.2	13.44	29.10	19.46
Intron-Sahara V2	**15.24**	12.02	**18.51**	**15.25**
GPT-4o Transcribe	28.54	15.03	30.80	24.79
Google Gemini-2.0 Flash	31.13	12.02	27.59	23.58
AWS Transcribe	42.22	14.02	30.08	28.77
Azure Speech Recognition	32.90	13.29	26.17	24.12
Google Chirp V3	39.6	17.53	38.57	31.9
Average	34.59	39.51	53.83	42.89

Table 4: Word Error Rate (WER %) for each model on the medical domain subsets of AfriSpeech-MultiBench, including clinical notes, medical dialogues, and doctor–patient conversations. Dataset full name mappings: Afri-Med: AfriSpeech Medical; Diag: AfriSpeech-Dialogue; Med.Conv: Med-Conv-Nig.

Model	**Name**	**Commands**	**Nums**
Parakeet-tdt-0.6B-v2	65.55	32.65	22.57
Parakeet-tdt-1.1B	76.44	33.67	26.47
Parakeet-rnnt-1.1B	75.78	35.36	26.66
Canary-1B-flash	75.69	30.05	20.15
Whisper-large-v3	73.1	31.58	18.11
Distil-Whisper-v3.5	68.15	37.28	15.28
CrisperWhisper	70.14	70.35	71.18
IBM Granite-3.3-2B	78.97	49.03	23.45
Voxtral (Mistral)	69.17	41.77	25.42
Canary-Qwen-2.5B	69.79	31.44	20.15
Phi-4 MM-Instruct	78.09	104.13	51.28
Intron-Sahara	24.24	**1.81**	14.81
Intron-Sahara V2	**12.4**	7.92	**4.27**
GPT-4o Transcribe	67.43	46.67	17.45
Google Gemini-2.0-Fl	74.12	40.77	18.23
AWS Transcribe	60.07	27.60	20.21
Azure	67.15	23.42	22.43
Google Chirp V3	84.91	40.22	24.83

Table 5: Word Error Rate (WER %) for each model on African named entites and Financial domain subsets of AfriSpeech-MultiBench. Dashes represent results that were not available.

5.4 Model size vs performance

Figure 3 and Table 3 show that, in a handful of domains, larger Speech LLMs (Granite, Phi-4, Voxtral, Canary-Qwen) only marginally outperform smaller architectures like conformer and Whisper variants half their sizes. In conversational speech, they are worse overall. Figure 3 indicates overall worse performance for open models with increasing size.

6 Discussion

This study yields a number of key insights that illuminate performance gaps and opportunities for advancing ASR systems in African settings:

6.1 Global benchmarks misrepresent African realities.

Leading models like Whisper and Parakeet achieve WERs below 10% on LibriSpeech and GigaSpeech, yet degrade to over 20-40% on African-accented data in AfriSpeech-MultiBench. This mismatch underscores the limits of current leaderboards in guiding ASR adoption across low-resource geographies.

6.2 Accent diversity drives large performance variance.

While models performed well on Kenyan and Ugandan English (average WERs as low as 12-18%), WERs doubled or tripled for West African and North African accents-exceeding 25% for many systems. This highlights the phonetic and prosodic diversity across the continent and the inadequacy of accent-agnostic training.

6.3 Conversational speech remains a major bottleneck.

Compared to read speech, performance worsened significantly on conversational corpora AfriSpeech-Dialog Medical, Medical Conversations (Med Convo), and Parliamentary speech. These mirror Western benchmarks, where models also struggle on AMI and Earnings22 relative to LibriSpeech or SPGISpeech. However, the drop-off in African conversational domains is more severe, revealing compound challenges likely due to accent, prosody, and domain shift.

6.4 Named entities and structured commands still confound models.

Most models scored above 40% WER on the AfriNames dataset, numbers, and financial voice commands, often failing to distinguish culturally unique or phonetically similar terms. This raises usability concerns in domains requiring accurate name capture or transactional integrity.

6.5 Model size and architecture don't predict reliability.

Smaller models like Parakeet-tdt-0.6B and Distil-Whisper sometimes matched larger peers on global benchmarks but showed inconsistent gains on African test sets. By contrast, Sahara a regionally optimized model consistently delivered best-in-class results across medical, legal, and conversational tasks.

Model	Nigeria	Ghana	Kenya	Rwanda	Uganda	South Africa	North Africa	Average
Parakeet-tdt-0.6B-v2	32.60	26.27	21.78	23.92	**9.38**	21.08	22.76	22.54
Parakeet-tdt-1.1B	28.65	22.08	21.21	20.39	21.35	25.96	24.13	23.40
Parakeet-rnnt-1.1B	32.76	23.69	23.19	24.71	23.08	27.59	24.42	25.63
Canary-1B-flash	29.00	25.06	18.25	20.39	23.65	26.30	24.04	23.81
Whisper-large-v3	26.53	27.22	16.70	21.18	16.16	19.85	20.72	21.19
Distil-Whisper-v3.5	26.69	22.16	18.51	23.53	19.22	24.45	21.43	22.28
CrisperWhisper	74.50	80.99	74.72	40.00	58.29	72.11	62.64	66.18
IBM Granite-3.3-2B	33.05	25.27	21.33	21.18	21.55	26.16	30.25	25.54
Voxtral (Mistral)	27.84	22.49	21.50	23.53	17.57	26.96	20.17	22.87
Canary-Qwen-2.5B	26.82	22.10	18.49	21.57	21.45	25.19	20.99	22.37
Phi-4 MM-Instruct	27.73	18.86	27.92	17.25	18.49	51.26	20.03	25.93
Intron-Sahara	15.85	**15.93**	**12.48**	18.04	12.26	**17.65**	**13.14**	15.05
Intron-Sahara V2	**12.83**	12.02	13.01	18.09	11.66	13.45	15.98	**13.86**
GPT-4o Transcribe	40.03	60.41	58.38	**15.69**	15.22	60.43	22.40	38.94
Google Gemini-2.0 Flash	26.47	24.54	21.29	21.57	19.61	27.59	22.11	23.31
AWS Transcribe	28.16	22.59	18.18	30.98	24.11	25.05	22.23	24.47
Azure Speech Recognition	26.41	23.01	17.59	25.49	19.31	25.10	25.46	23.20
Average	30.99	27.54	24.36	23.01	21.25	29.65	24.38	25.31

Table 6: Word Error Rate (WER %) for each model across African accents in AfriSpeech-MultiBench, extended to include Voxtral (Mistral) and Intron-Sahara V2.

6.6 Benchmarking must evolve beyond average-case accuracy.

AfriSpeech-MultiBench enables fine-grained, domain-aware evaluation that reflects real-world deployment conditions. It provides not only model ranking, but also insight into where and why systems fail offering practical guidance for building domain and region-specific ASR solutions in healthcare, law, finance, and public service delivery across Africa.

7 Conclusion

This study set out to address the gap between global ASR benchmarks and real-world performance on African-accented, domain-specific speech. Through AfriSpeech-MultiBench, we reveal that top-performing models on standard datasets like LibriSpeech and TED-3 achieving sub-5% WER can exhibit 5-10X higher error rates on African speech, especially in medical, financial, and conversational domains. These disparities are consistent across open-source and proprietary systems, highlighting persistent geographic, linguistic, and domain biases in existing ASR development and evaluation pipelines.

Our findings underscore the need for regionally grounded benchmarks and models. Intron-Sahara, a model trained with African-specific data, consistently outperformed global leaders across domains and accents, particularly in name recognition, doctor patient dialogue, and financial commands. By benchmarking 18 models across 8 African countries and 7 key domains, AfriSpeech-MultiBench provides actionable insights for building inclusive ASR systems. This work lays the foundation for future research and deployment efforts in healthcare, legal transcription, customer service, and multilingual voice applications across the African continent.

Limitations

While AfriSpeech-MultiBench offers a broad and diverse benchmark across African-accented English, several limitations warrant consideration. First, despite including over 10 countries and six domains, the benchmark does not yet cover all major linguistic regions in Africa or fully represent under-resourced countries with limited public data availability. Certain domains such as manufacturing, education, and public safety are not currently included, and even within included sectors like healthcare and finance, dataset sizes remain modest compared to global corpora, which may limit fine-grained error analysis and generalization of results.

Secondly, some datasets used are proxies rather than fully representative of their target verticals. For instance, parliamentary proceedings may not fully capture the legal domain's complexity, such as courtroom vernacular, legalese, or multilingual code-switching common in legal aid and judicial settings. Similarly, due to privacy constraints, customer support datasets from private call centers were not included, limiting direct benchmarking for commercial deployments. These gaps highlight both the urgent need and the opportunity for continued investment in domain-specific and geographically expansive data collection to build more comprehensive benchmarks for inclusive speech technologies.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba Oluwadara Alabi, Xuanli He, and 1 others. 2025. Irokobench: A benchmark for african languages in the age of large language models. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 2732–2757, Mexico City, Mexico. Association for Computational Linguistics.

Galen Andrew, Mingqing Chen, Jinyu Lu, and Kevin Sim. 2022. Earnings22: A 100-hour benchmark corpus for earnings-call ASR. In *Proceedings of Interspeech 2022*, pages 3158–3162.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, and 1 others. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 4218–4222.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 12449–12460.

Jean Carletta, Simone Ashby, Séverine Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, and 1 others. 2005. The AMI meeting corpus: A pre-announcement. In *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction (MLMI)*, pages 30–44. Springer.

Chang Chen, Yiming Peng, Yuan Guo, Nanxin Yang, Shuai Zhang, Yongqiang Cui, and 1 others. 2021. Gigaspeech: An evolving, multi-domain ASR training corpus with 10,000 hours of audio. In *Proceedings of Interspeech 2021*, pages 3790–3794.

Zhehuai Chen, He Huang, Andrei Andrusenko, Oleksii Hrinchuk, Krishna Puvvada, Jason Li, and 1 others. 2023. SALM: Speech-augmented language model with in-context learning for speech recognition and translation. *arXiv preprint arXiv:2310.09424*.

Alex DiChristofano, Henry Shuster, Shefali Chandra, and Neal Patwari. 2022. Global performance disparities between english-language accents in automatic speech recognition. In *arXiv preprint arXiv:2208.01157*.

P. Dossou, Bonaventure F.2025. Advancing african-accented english speech recognition: Epistemic uncertainty-driven data selection for generalizable ASR models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, Bangkok, Thailand. Association for Computational Linguistics.

Daniel Galvez, Vladimir Bataev, Hainan Xu, and Tim Kaldewey. 2024. Speed of light exact greedy decoding for rnn-t speech recognition models on gpu. In *Interspeech 2024*, pages 277–281.

Cong Guo, Jing Zhang, Xiaohui Ma, Yongqiang Huang, Mike Lewis, Zhe Wei, and Shiliang Chen. 2022. SPGISpeech: 5,000 hours of transcribed financial audio for self-supervised speech representation learning. In *Proceedings of Interspeech 2022*, pages 3663–3667.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève. 2018. TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *Speech and Computer (SPECOM)*, pages 198–208.

Intron Health. 2025. Afrispeech-countries: Cross-regional african-accented english speech benchmark. https://huggingface.co/datasets/intronhealth/afrispeech-countries.

Intron Health. 2025a. Afrispeech-parliament: Transcribed parliamentary sessions from four african nations. https://huggingface.co/datasets/intronhealth/afrispeech-parliament.

Intron Health. 2025b. Afri-names: African named-entity read-speech corpus. https://huggingface.co/datasets/intronhealth/afri-names.

Intron Health. 2025c. Med-convo-nig: Nigerian doctor–patient tele-consultation speech dataset. https://huggingface.co/datasets/intronhealth/med-convo-nig.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, and 1 others. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

Alexander H. Liu, Andy Ehrenberg, Andy Lo, Clément Denoix, Corentin Barreau, Guillaume Lample, Jean-Malo Delignon, Khyathi Raghavi Chandu, Patrick von Platen, Pavankumar Reddy Muddireddy, Sanchit Gandhi, Soham Ghosh, Srijan Mishra, Thomas Foubert, Abhinav Rastogi, Adam Yang, Albert Q. Jiang, Alexandre Sablayrolles, Amélie Héliou, and 87 others. 2025. Voxtral. *Preprint*, arXiv:2507.13264.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations:

What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, Sahib Singh, Bonaventure F. P. Dossou, and 1 others. 2023. Afrispeech-200: Pan-african accented speech dataset for clinical and general domain asr. *Transactions of the Association for Computational Linguistics*, 11:1599–1617.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Krishna C. Puvvada, Piotr Żelasko, He Huang, Oleksii Hrinchuk, Nithin Rao Koluguri, Kunal Dhawan, Somshubra Majumdar, Elena Rastorgueva, Zhehuai Chen, Vitaly Lavrukhin, Jagadeesh Balam, and Boris Ginsburg. 2024. Less is more: Accurate speech recognition & translation without web-scale data. *arXiv preprint arXiv:2406.19674*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Dima Rekesh, Nithin Rao Koluguri, Samuel Kriman, Somshubra Majumdar, Vahid Noroozi, He Huang, Oleksii Hrinchuk, Krishna C. Puvvada, Ankur Kumar, Jagadeesh Balam, and Boris Ginsburg. 2023. Fast conformer with linearly scalable attention for efficient speech recognition. *arXiv preprint arXiv:2305.05084*.

Hadeel Saadany, Catherine Breslin, Constantin Orasan, and Sophie Walker. 2023. Better transcription of uk supreme court hearings. In *AI4AJ@ICAIL*.

Mardhiyah Sanni, Tassallah Abdullahi, Devendra Kayande, Emmanuel Ayodele, Naome Etori, Michael Mollel, and 1 others. 2025. Afrispeech-dialog: A benchmark dataset for spontaneous english conversations in healthcare and beyond. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Holger Schwenk, Loïc Barrault, Yu-An Chung, Francisco Guzmán, Juan Pino, and the Seamless Communication Team. 2023. Seamlessm4t: Massively multilingual and multimodal machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Marieke M. van Buchem, Hileen Boosman, Martijn P. Bauer, Ilse M. J. Kant, Simone A. Cammel, and Ewout W. Steyerberg. 2021. The digital scribe in clinical practice: A scoping review and research agenda. *npj Digital Medicine*, 4:57.

Lingli Wang, Ni Huang, Yili Hong, Luning Liu, Xunhua Guo, and Guoqing Chen. 2023. Voice-based AI in call center customer service: A natural field experiment. *Production and Operations Management*, 32(4):1002–1018.

Weiyi Wang, Chau Tran, Fahim Azhar, Henrik Rottmann, Armand Joulin, and 1 others. 2021. Voxpopuli: A large-scale multilingual speech corpus for representation, semi-, and self-supervised learning. In *Proceedings of Interspeech 2021*, pages 993–997.

Mario Zusag, Laurin Wagner, and Bernhad Thallinger. 2024. Crisperwhisper: Accurate timestamps on verbatim speech transcriptions. In *Interspeech 2024*, pages 1265–1269.

Appendix

Noise Robustness

Model	**SIL**	**Short**	**Int.SIL**	**Avg.**
Parakeet-tdt-0.6B-v2	73.12	34.50	15.05	40.89
Parakeet-tdt-1.1B	78.31	33.20	22.35	44.62
Parakeet-rnnt-1.1B	211.62	35.41	23.86	90.30
Canary-1B-flash	71.02	54.56	30.39	51.32
Whisper-large-v3	41.90	36.58	22.89	33.79
Distil-Whisper-v3.5	43.31	37.38	23.28	34.00
CrisperWhisper	53.62	36.07	24.76	38.82
IBM Granite-3.3-2B	43.90	60.75	32.92	45.86
Voxtral (Mistral)	40.77	57.57	32.18	43.51
Canary-Qwen-2.5B	37.63	54.39	31.43	41.15
Phi-4 MM-Instruct	154.05	136.60	75.71	122.45
Intron-Sahara	25.49	21.47	9.77	18.91
Intron-Sahara V2	**0.00**	**15.98**	**7.59**	**7.86**
GPT-4o Transcribe	44.48	37.80	18.48	33.59
Google Gemini-2.0 Flash	45.45	37.80	18.48	33.91
AWS Transcribe	43.45	35.36	21.12	33.98
Azure Speech Recog.	44.27	32.52	21.04	32.61
Google Chirp V3	31.37	40.62	23.10	31.70
Average	58.86	45.62	26.18	43.55

Table 7: Robustness evaluation across `SIL`, `Short`, and `Int.SIL` subsets. The final column shows their mean (`Avg.`).

Pre- and Post-Processing

Audio pre-processing

Audio files are used exactly as distributed by the source datasets; no further segmentation or concatenation is performed. A single exception concerns the NVIDIA NeMo checkpoints (*parakeet-**, *canary-1B*), which require 16kHz mono input. When a file is multi-channel or sampled above 16kHz, it is down-mixed to mono and re-sampled with `sox` prior to inference. All other engines (Whisper variants, API endpoints) accept the original wave-forms without modification.

Transcript pre-processing

Reference and hypothesis strings undergo a three-stage normalisation pipeline, implemented exactly as in the public evaluation script:

1. `clean_text` — lower-cases, trims whitespace, removes punctuation, deletes 32 variants of *[inaudible]*, and removes frequent filler words (*uh*, *hmm*,*mmhmm*,...).
2. `text_to_numbers` — maps number words (*"twenty"* → 20) and ordinal words (*"first"* → 1st) to their digit form.
3. `EnglishTextNormalizer` — applies the Whisper normaliser for final case-folding and whitespace cleanup.

A sentinel token `abcxyz` replaces empty strings to avoid undefined denominators in word-error calculations.

Post-processing for Nemo models

NeMo/Parakeet outputs include automatically generated punctuation. Before the three-stage normaliser, inverse text normalisation is applied to restore standard spacing around commas and periods, ensuring a fair comparison with punctuation-free reference strings.

Metric

Word-error rate (WER) is computed with JIWER

$$\mathrm{WER}(r, h) = \frac{S + D + I}{|r|},$$

where S, D and I count substitutions, deletions and insertions needed to transform hypothesis h into reference r.

Prompting for Speech Augmented Language Models

Default prompts for open source speech augmented language models where used:

- Canary-Qwen-2.5B : "Transcribe the following: model.audio_locator_tag", "audio": ["speech.wav"]
- Mixtral (Voxtral-Mini-3B-2507): We used its apply_transcription_request function which takes an audio file and wraps it with inbuilt prompts for speech transcription.
- Google Gemini 2.0 Flash: The required prompt according to Google API documentation was used, prompt = """ Transcribe this ENGLISH audio. """
- Phi-4 Multimodal Instruct: <|user|><|audio_1|>Transcribe the audio to text<|end|><|assistant|>