Voice of a Continent: Mapping Africa's Speech Technology Frontier

AbdelRahim Elmadany[ξ] **Sang Yun Kwon**[ξ] **Hawau Olamide Toyin**[Ω]
Alcides Alcoba Inciarte[ξ] **Hanan Aldarmaki**[Ω] **Muhammad Abdul-Mageed**[ξ,λ]
[ξ]The University of British Columbia [Ω]MBZUAI [λ] Invertible AI
{a.elmadany,muhammad.mageed}@ubc.ca

Abstract

Africa's rich linguistic diversity remains significantly underrepresented in speech technologies, creating barriers to digital inclusion. To alleviate this challenge, we systematically map the continent's speech space of datasets and technologies, leading to a new comprehensive benchmark *SimbaBench* for downstream African speech tasks. Using *SimbaBench*, we introduce the *Simba* family of models,[1] achieving state-of-the-art performance across multiple African languages and speech tasks. Our benchmark analysis reveals critical patterns in resource availability, while our model evaluation demonstrates how dataset quality, domain diversity, and language family relationships influence performance across languages. Our work highlights the need for expanded speech technology resources that better reflect Africa's linguistic diversity and provides a solid foundation for future research and development efforts toward more inclusive speech technologies.

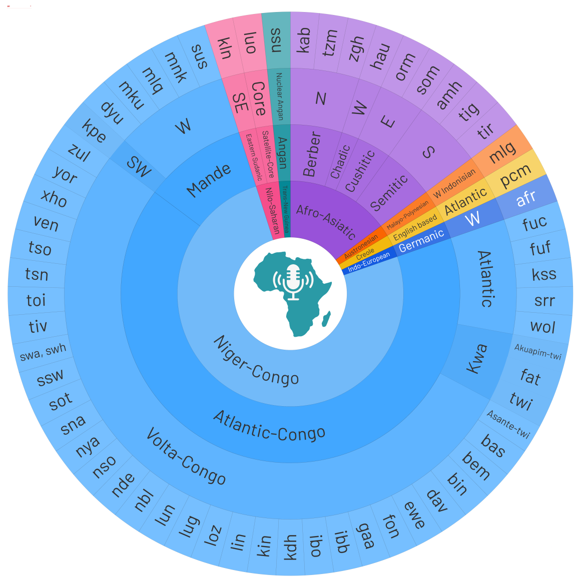


Figure 1: A three-level language family hierarchy illustrating the 61 African languages included in our analysis, benchmark, and speech modeling efforts.

1 Introduction

Speech is one of the most natural and fundamental forms of human communication. Advances in speech technologies, such as automatic speech recognition (ASR), text-to-speech (TTS), and spoken language understanding, have enabled transformative applications including virtual assistants, real-time translation, and accessible communication tools for people with disabilities. However, the benefits of these technologies are not equitably distributed. Most current resources and research efforts are concentrated on a handful of widely spoken languages, particularly English, leaving the majority of the world's linguistic diversity underrepresented (Bender, 2011; Joshi et al., 2020). This imbalance is especially stark in the context of African languages, which are spoken by hundreds of millions but often lack the data and tools necessary for the development of robust speech systems. Addressing this gap is crucial for fostering technological inclusion, preserving linguistic heritage, and enabling culturally relevant digital innovation. Moreover, as large language models (LLMs) increasingly integrate speech capabilities (Huang et al., 2024; Cui et al., 2024; Nguyen et al., 2023, 2025), ensuring that African languages are supported in both text and speech modalities is essential for equitable access to emerging AI technologies.

While recent multilingual speech models such as Whisper (Radford et al., 2022), MMS (Pratap et al., 2023), and SeamlessM4T (Anastasopoulos et al., 2023) include some coverage of African languages, their performance on key speech tasks such as ASR, TTS, and spoken language identification (SLID) remains inadequate, especially for low-resource and tonal languages (Alabi et al., 2024; Hyman, 2003). Despite recent efforts to improve speech model-

[1]https://github.com/UBC-NLP/simba

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 11028–11050
November 4-9, 2025 ©2025 Association for Computational Linguistics

ing for African languages like mHuBERT (Zanon Boito et al., 2024) and AfriHUBERT (Alabi et al., 2024), these models cover only a small fraction of Africa's languages. In addition, African speech datasets are often undocumented or fragmented, with little clarity on their scope, supported tasks, language coverage, and evaluation standards.

Recognizing the critical need to clearly characterize the current landscape of African speech datasets and technologies, we undertake a mapping of these resources and systems. In particular, we offer a number of contributions: **(1) New Speech Benchmark:** we conduct extensive data collection and aggregate and harmonize all publicly available resources covering ASR, TTS, and SLID tasks. This dataset collection spans diverse linguistic families and geographic regions, leading the way to the development of ***SimbaBench***, a unified benchmark designed specifically for African speech processing. **(2) Data-Driven Coverage Analysis:** with *SimbaBench* at hand, we carry out a quantitative mapping of current speech datasets in Africa, allowing us to draw connections between dataset availability across languages and populations. This helps paint the picture for the current state of African speech resources. **(3) Model Evaluation:** we benchmark existing state-of-the-art (SoTA)[2] African and multilingual speech models on *SimbaBench*, thereby empirically assessing capabilities and limitations of these models across African speech tasks. These evaluations offer critical insights into where current models fall short and where targeted innovation is needed. **(4) A Family of SoTA African Speech Models:** we exploit our datasets to build upon existing models, introducing a suite of fine-tuned models, dubbed *Simba*, achieving SoTA performance on a wide set of African languages across the downstream tasks. Figure 2 illustrates the methodological workflow employed in our work.

Through this work, we provide foundational tools (i.e., *SimbaBench* and *Simba* models) and resources to accelerate speech technology for African languages and invite community participation in this inclusive, multilingual effort. The paper is organized as follows: Section 2 overviews related work in African NLP and Speech. Section 3 describes our data mapping and collection process.

[2]We use 'SoTA' to refer to best performance achieved among all systems evaluated under *SimbaBench's* unified benchmarking conditions, establishing a reproducible baseline for future research.

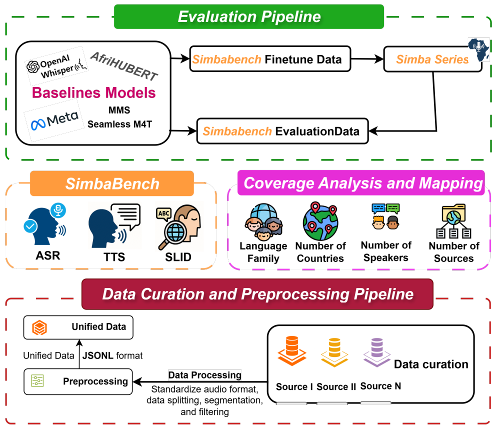


Figure 2: Methodological workflow, illustrating the three main components: (1) the data curation and preprocessing pipeline, (2) *SimbaBench* with quantitative mapping of current speech datasets in Africa, and (3) the evaluation pipeline.

We detail our benchmark *SimbaBench* in Section 4. Section 5 outlines our evaluation setup, and we discuss results and findings in Sections 6 and 7.

2 Literature Review

Speech and language technologies enable broader access to information and can potentially support and promote linguistic diversity. However, of the over 7,000 languages spoken worldwide, only a select few are represented in contemporary language technologies and applications (Joshi et al., 2020). Most speech and NLP systems are predominantly trained on a limited subset of languages, primarily from dominant language families and specific geographies, leaving most languages unrepresented (Ponti et al., 2019). Joshi et al. (2020) categorizes languages into 6 classes based on available resources, ranging from ***The Left-Behinds*** (Class 0) with virtually no digital presence to ***The Winners*** (Class 5) with abundant resources and technological support. With most African languages occupying the lower tiers of this classification, a substantial language gap persists, leaving indigenous and regional languages under-represented in NLP (Adebara et al., 2025), highlighting the need for additional efforts to promote inclusive language technologies (Ojo et al., 2023).

Progress in African NLP. In recent years, significant progress has been made towards improving representation and performance of African languages in NLP, particularly in text understanding and generation tasks (Adebara and Abdul-Mageed, 2022; Adebara et al., 2025). Benchmarks like SAHARA (Adebara et al., 2025), IrokoBench (Adelani et al., 2024), and others (Ojo et al., 2023; Wang et al., 2023; Oladipo et al., 2023; Reid et al., 2021) have advanced NLU and NLG capabilities. In terms of model development, models like AfroXLMR (Belay et al., 2025), Cheetah (Adebara et al., 2024) and others have contributed significantly to these developments (Adebara et al., 2022b; Elmadany et al., 2024; Adebara et al., 2022a). Despite these advancements, the development of African speech technologies remains slow, impeded by intensive computational requirements, a shortage of large-scale speech corpora, and historical bias towards high-resource Western languages (Joshi et al., 2020). This resource gap motivates our work toward inclusive technologies for Africa's diverse languages.

Progress in African Speech. Prior work on African speech has focused primarily on speech resource collection and the presentation of baseline results (Ogun et al., 2024; Gutkin et al., 2020; Sikasote et al., 2023; Meyer et al., 2022). Projects such as the CMU Wilderness dataset provided early bootstrapped text-to-speech voices for over 100 African languages (Black, 2019), laying essential groundwork. This has been further advanced by structured community initiatives such as the BULB Project, which focused on breaking barriers for unwritten languages (Adda et al., 2016), and the AI4D African Language Program, designed to foster the creation of more comprehensive speech resources (Siminyu et al., 2021). Although large-scale speech models are becoming increasingly multilingual, most African languages remain underrepresented, having been excluded during the pretraining of major speech models (Alabi et al., 2024). Despite these limitations, recent efforts such as AfriHuBERT (Alabi et al., 2024) exemplify progress toward addressing this gap. Moreover, multilingual speech recognition for African languages using self-supervised learning has been demonstrated (Ritchie et al., 2022), complementing academic work on Africa-centric, self-supervised pre-trained models for multilingual speech representation in a Sub-Saharan context (Caubrière and Gauthier, 2024).

3 Mapping the Data Landscape

To understand the current state of speech technology for African languages, we begin with a comprehensive assessment of the available data resources. This is a necessary step before evaluating the capabilities of current models or proposing new directions for building robust multilingual systems. In particular, it is essential to identify what resources are available, where they originate, and where critical gaps persist. Below, we present an overview of publicly available African speech corpora, encompassing both labeled and unlabeled audio data. Our analysis centers on three core speech downstream tasks: ASR, TTS, and SLID.

Our objective is beyond mere data collection. Rather, our aim is to map the linguistic and acoustic diversity represented within existing datasets. This mapping lays the groundwork for a comprehensive and inclusive data infrastructure that authentically represents the multilingual realities of the African continent, which, as emphasized by Adebara et al. (2025), is crucial for ensuring equitable participation in global language technology advancements.

Task	Dataset	#Lang.	Dur. (h)	Domain
ASR	Alffa Public (Besacier and Gauthier, 2023)	4	58.66	RS, N
	BembaSpeech (Sikasote and Anastasopoulos, 2022)	1	26.93	N, V
	Common Voice (CV-19) (Mozilla Foundation, 2023)	21	1,843.65	RS
	Financial Speech (Asamoah Owusu et al., 2022)	4	149.55	RS, F
	Kallaama (Gauthier et al., 2024)	3	113.68	R, IR
	Lwazi (Van Heerden et al., 2016)	10	42.80	TC
	Naija Voices (NaijaVoices, 2024)	3	1,867.52	RS
	NCHLT + AUX1/2 (Barnard et al., 2014)	11	1,922.05	RS
	Nicolingua (0004) (Doumbouya et al., 2021)	3	1.24	R
	YorubaVoice (Gutkin et al., 2020)	1	4.03	G
	Zambezi Voice (ASR) (Sikasote et al., 2023)	3	54.23	RS, TS
	SO (Code-Switched) (der westhuizen and Niesler, 2018)	4	14.27	TV
	SPCS (Code-Switched) (Modipa et al., 2015)	1	10.48	R
	ASR Statistics	**42**	**6,109.09**	
SLID	Nicolingua (0003) (Doumbouya et al., 2021)	6	143.75	R
	OlongoAfrica (Ours)	10	2.40	SS
	UDHR (Ours)	6	1.05	HR
	Voice of Africa (VOA) (Ours)	10	865.08	N
	VoxLingua (Valk and Alumäe, 2021)	9	773.66	V
	Zambezi Voice (Audio Only) (Sikasote et al., 2023)	5	176.00	TS
	SLID Statistics	**39**	**1,961.94**	
TTS	BibleTTS (Meyer et al., 2022)	6	306.69	RB
	High-Quality TTS (SA) (van Niekerk et al., 2017)	4	13.16	WS
	Kinyarwanda TTS (Digital Umuganda, 2023)	1	14.08	—
	TTS Statistics	**11**	**333.93**	
	AfriSpeech (Accented-African)) (Olatunji et al., 2023)	1	200	C, G
	Overall	**61**	**8,604.96**	—

Table 1: Overview of curated African audio datasets used in our data. This summary includes dataset type, number of languages covered (*#Lang.*), total duration in hours (*Dur.*), and source domain. "Ours" refer to new data that we primarily collected or curated as part of this work. **RS.** refers to Read Speech, **TS.** Talk Show **TC.** Telephone Conversations, **F.** Financial, **TV.** TV Shows, **IR.** Interviews, **N.** News, **C.** Clinical, **SS.** Short Stories, **G.** General, **HR.** Human Rights, **R.** Radio, **V.** Video, **SO.** Soap Opera, **WS.** Wikipedia-based Speech, and **RB** Read Bible.

3.1 Data Curation

We curate a large-scale corpus of publicly available audio data integrating both labeled and unlabeled speech to ensure broad linguistic, acoustic, and demographic coverage. In total, we aggregate $8,605$ hours of audio drawn from 26 publicly available sources, comprising well-established corpora for downstream tasks (supervised) as well as large-scale unlabeled speech data (unsupervised). The collected resources span multiple domains, including media-rich and culturally grounded sources. This introduces variability in speech styles, regional dialects, and speaker identities—dimensions often underrepresented in traditional benchmarks.

African Data. We collect over $8,380$ hours of clean data spanning 61 african languages. Consisting of richly diverse domains like, *broadcast*, *radio*, *read speech*, and *spontaneous conversations*. This includes $6,080$ hours of ASR covering 42 languages, 334 hours of TTS spanning 11 languages, and $1,960$ hours of untranscribed (audio-only) data across 32 languages for SLID.

Code-switched Data. We include ~ 34 hours of code-switched speech data, encompassing seven language pairs that combine African and non-African languages within a single utterance. These recordings reflect authentic patterns of multilingual discourse in everyday African contexts and are essential for training models capable of handling spontaneous, mixed-language input.

African-accented English. Furthermore, we incorporate 200 hours of African-accented English speech, representing 120 distinct accents from 13 African countries with $2,463$ unique speakers (Olatunji et al., 2023).

A comprehensive summary of the dataset composition, language distribution, and task coverage is presented in Table 1, with additional details provided in Appendix §A. Also, Table B.1 (in Appendix §B) provides detailed information on the total audio duration (in hours) for each language across various datasets.

3.2 Data Preprocessing and Standardization

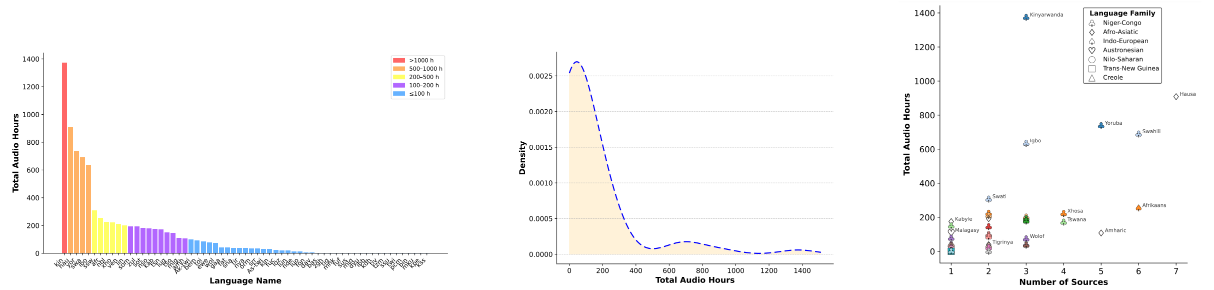
To ensure consistency, quality, and usability across the diverse audio datasets, we apply a unified preprocessing pipeline encompassing *format standardization*; converting all audio to 16 kHz mono WAV format, *segmentation*, *filtering*, and *noise removal*; breaking long recordings into 1-20 second utterances and eliminating excessive noise, and *metadata consolidation*; reformatting datasets into a unified JSON schema with standardized fields. Our preprocessing pipeline enables robust training and evaluation across diverse African speech corpora, establishing a foundation for consistent benchmarking and inclusive model development across all downstream tasks. More detailed information is outlined in Appendix §C.

3.3 Quantitative Data Analysis

We present a quantitative analysis of the curated audio resources with respect to language distribution, task-specific coverage, and overall data volume. Our findings reveal disparities in speech data availability across African languages. We highlight the strengths and limitations of current African speech datasets, informing the feasibility of training and evaluating models for the aforementioned tasks. Together, these trends underscore the need for strategic data collection that prioritizes not only volume but also domain diversity and equitable representation across linguistic and demographic factors. Without such targeted efforts, existing disparities for African speech technology development will likely persist or worsen, further marginalizing already under-resourced languages.

Overall Data Distribution. Africa has more than 2000 languages and dialects, of which our extensive efforts could only identify 61 that have publicly available data. Even within this small number of languages, we find a minority of languages—*Kinyarwanda*, *Hausa*, *Yoruba*, *Swahili*, and *Igbo*—accounting for hundreds to thousands of hours of recorded speech, whereas the majority of languages have fewer than one hour of data as shown in Figure 3a. Figure 3b highlights this imbalance through the Kernel Density Estimate (KDN) of total hours collected, revealing a heavily right-skewed distribution with high density concentrated near zero hours and a long tail extending toward the few resource-rich languages. This pattern highlights the imbalance driven primarily by targeted collection efforts rather than linguistic or demographic representation.

Language Family Distribution. The distribution by language family in Figure 1 shows that the ***Niger-Congo*** and ***Afro-Asiatic*** families dominate the available resources. Within the ***Niger-Congo*** group, *Kinyarwanda*, *Yoruba*, *Igbo*, *Swahili*, and several *Volta-Congo* languages account for the largest volumes of data. From the ***Afro-Asiatic***



(a) Audio hours per language.

(b) Distribution density (KDE).

(c) Hours vs. number of sources.

Figure 3: Speech data distribution across the 61 African languages in collected data, highlighting volume, density, and source diversity.

family, *Hausa* possesses substantial resources, whereas *Somali*, *Amharic*, and *Tamazight* remain comparatively under-represented. Other families like ***Nilo-Saharan***, ***Austronesian***, and ***Trans-New Guinea*** appear only sparsely, with *Malagasy* as the primary exception within the ***Austronesian*** family. Overall, the majority of languages are from the ***Niger-Congo*** and ***Afro-Asiatic*** families, reflecting the dominant language groups in Africa.

Native Speaker Distribution. We find a clear mismatch between speaker population size and available audio resources. Languages with large speaker populations often have minimal data—for example, *Oromo*, with 45M speakers, has only 34 hours of audio, while *Nigerian Pidgin*, spoken by roughly 120M people, has just 0.21 hours. In contrast, some languages with smaller populations are comparatively well-resourced, such as *South Ndebele* (2.4M speakers, 223 hours) and *Swati* (4.7M speakers, 307 hours). These disparities suggest that data availability correlates more strongly with a number of potential factors such as language use in media, data archiving and accessibility, and targeted collection initiatives than with population size. Collectively, these factors are directly related to adopted language policies (Adebara et al., 2025).

Number of Sources. The number of data sources per language indicates that overall volume is primarily driven by inclusion in major collection projects rather than by a broad diversity of smaller efforts. Figure 3c illustrates this relationship between source diversity and total hours collected. High-volume languages either appear in multiple major sources—such as *Hausa* (7 sources) and *Swahili* (6 sources)—or derive substantial coverage from a single extensive initiative, as with *Kinyarwanda* via CV-19 and *Igbo* via NaijaVoice. In contrast, lower-resource languages are typically represented only through isolated small-scale efforts. Overall, data volume is dictated more by the scale of one or two dominant collections than by the sheer number of sources. A single large dataset can secure extensive hours but often limited in domain diversity, whereas multiple smaller sources may yield less total audio yet provide broader, more balanced coverage for downstream speech applications.

Dataset Fragmentation. Our analysis reveals that dataset fragmentation represents a significant barrier to reproducible research in the African speech data landscape. This challenge manifests as a lack of standardized training and testing splits across many key corpora,[3] preventing fair comparisons between studies. Furthermore, many datasets exhibit severe imbalance between large portions of unlabeled audio and minimal labeled sets, limiting their utility on supervised tasks.[4] These inconsistencies underscore the critical need for a unified benchmark to standardize evaluation and unlock the full potential of these valuable but fragmented resources.

4 *SimbaBench* Benchmark

Motivation. To address the lack of standardized benchmarks for African speech technologies, we introduce *SimbaBench* —a unified evaluation suite designed to support diverse African speech tasks. It enables consistent model assessment, fosters reproducible research, and promotes fair comparisons, advancing inclusive language technologies for underrepresented communities.

[3]Examples are the Lwazi (Van Heerden et al., 2016) and CS Soap Opera (der westhuizen and Niesler, 2018) datasets.

[4]A clear example is the Nicolingua-0003 corpus (Doumbouya et al., 2021).

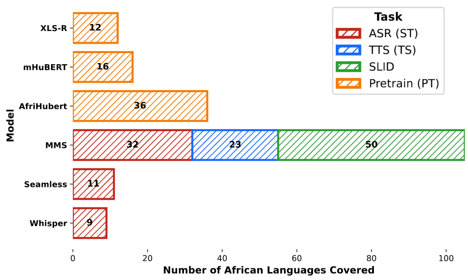


Figure 4: Comparison of African language coverage across the downstream tasks as well as pretraining.

Coverage. *SimbaBench* unifies all publicly available African speech datasets (Section 3), encompassing a wide range of languages, dialects, and domains. It supports comprehensive evaluation across both high- and low-resource languages through three core tasks: ASR, TTS, and SLID. Each task is paired with curated datasets and standardized metrics to enable consistent, fair comparisons across models and languages.

Data Splits and Release. To ensure consistency and reproducibility, we adopt official training and test splits when available; otherwise, we apply a 90%-10% train-test partition. For model development and checkpoint selection, we construct a multilingual training and development set by sampling n hours of training data (5 hours per language for ASR, 12 for TTS) and 30 minutes of development data per language when available. Evaluation is conducted per dataset to enable comparability with prior work and highlight dataset-specific challenges. We release the multilingual training and development splits to support benchmarking and tuning, while test sets are shared via standardized configuration files. *SimbaBench* will be hosted on the Hugging Face Datasets platform.[5]

5 Model Evaluation on *SimbaBench*

We evaluate *SimbaBench* on several leading *open*-source models to asses their generalization ability in the contexts of African languages and provide insights for future model development. Below we describe our evaluation pipeline in detail and the baseline models used for evaluation.

5.1 Baseline Models

We benchmark several state-of-the-art multilingual speech models with varying architectures and training approaches to assess their performance on African language audio data. We evaluate Whisper (Radford et al., 2022), Seamless (Anastasopoulos et al., 2023), MMS (Pratap et al., 2023), AfriHUBERT (Alabi et al., 2024), and Wav2Vec2-XLS-R (Babu et al., 2021).

Figure 4 illustrates the extent to which these baseline models cover African languages in their pretraining or supervised finetuning. The figure shows that MMS offers the broadest African language coverage across tasks, while models like AfriHUBERT provide the highest coverage in unsupervised pretraining. Whisper-v3 and SeamlessM4T-v2 provide limited ASR support, highlighting both task-specific strengths and existing gaps in African language inclusion. Table D.1(Appendix §D) presents a detailed overview of African language support across models for pretraining and various downstream tasks in speech and language processing. Collectively, these models establish strong baselines for evaluating the current state of ASR technology for African languages.

5.2 *Simba* Series

In addition to evaluating existing speech models as described above, we finetune a series of models, referred to as the *Simba* Series, leveraging the multilingual training and development sets from *SimbaBench* for the three downstream tasks. The *Simba* models are designed to enhance performance and mitigate language coverage gaps identified in prior baselines.

***Simba*-ASR.** We finetune five baseline models (see §5.1 for details) using the *SimbaBench* multilingual training and development sets.[6] This multilingual setup enables the development of five new ASR models, each adapted specifically to African linguistic contexts. The resulting models are *Simba*-H, finetuned from AfriHuBERT, *Simba*-M from MMS-1b-all, *Simba*-S from SeamlessM4T-v2-MT, *Simba*-X from Wav2Vec2-XLS-R, and *Simba*-W from Whisper-v3-large. All models are finetuned in a multilingual fashion. We follow the same protocols for multilingual training as described in the original Whisper, MMS, and Seamless models. For XLS-R, mHuBERT, and AfriHubert, we adopt a simple strategy of multilingual finetuning

[5] See project GitHub: https://github.com/UBC-NLP/simba.

[6] ASR finetuning data comprise 215 hours of transcribed training audio (5 hours per language) and 21.5 hours of validation audio (30 minutes per language), covering 43 African languages.

Language	Test Set	MMS	Seamless	Whisper	WhisperT	*Simba* Series (Ours)				
						Simba-H	*Simba*-M	*Simba*-S	*Simba*-X	*Simba*-W
Akuapim-twi (aka)	FS	85.82/40.14	219.67/190.49	1181.0/1131.23	499.51/547.24	26.83/10.13	17.6/8.13	13.29/8.45	23.74/10.35	29.1/19.1
Asante-twi (aka)	FS	83.6/32.35	230.88/196.71	665.34/574.27	245.5/222.37	26.78/7.36	13.87/5.38	7.06/2.62	19.93/7.06	15.63/7.98
Afrikaans (afr)	Lwazi	92.06/37.59	37.91/16.47	66.05/34.32	73.17/39.05	62.81/17.9	36.29/9.86	15.62/4.99	102.96/53.45	29.22/11.0
...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...
Zulu (zul)	Lwazi	70.12/32.66	107.96/84.77	164.54/106.64	78.11/43.35	62.92/17.57	38.58/10.88	108.53/103.61	101.93/52.87	27.63/10.87
Zulu (zul)	NCHTL	31.31/5.12	74.28/20.56	648.45/244.13	379.87/134.73	30.55/4.69	26.36/3.96	23.87/4.47	60.96/8.79	33.92/5.71
Overall Average		75.9/35.26	146.69/98.92	611.91/437.98	196.7/149.79	59.9/21.46	48.11/17.41	41.65/18.3	82.64/39.31	60.56/31.16

Table 2: Comparison of ASR performance across various African languages using baseline models and our *Simba* models. Evaluation metrics are reported as `WER/CER`. Red underlines indicate that the model does not support the corresponding language, while green highlights denote the best-performing model for each language or test set. Full results are provided in Table F.1 (Appendix §F).

by adding a CTC layer [7] and updating all parameters.

***Simba*-TTS.** As the only baseline model that supports TTS, we finetune the MMS-TTS model (Pratap et al., 2023) extending support to additional African languages. The original MMS-TTS model only supports 4 out of the 11 African languages included in our collection. As a result, unlike the ASR setup, we do not finetune on the entire multilingual dataset; instead, we focus exclusively on the 7 African languages previously not supported by MMS-TTS, and for which TTS data exist in our collection. For each language, we independently finetune from existing MMS-TTS checkpoints belonging to linguistically similar languages, selecting the best-performing checkpoint based on validation performance. Specifically, *Akuapem Twi* and *Asante Twi* are finetuned from the *Akan* checkpoint; *Tswana* and *Southern Sotho* from the *Tsonga* checkpoint; *Afrikaans* from the *Dutch-based creole* checkpoint, reflecting its linguistic history; and *Lingala* from the *Swahili* checkpoint. This language-family-based knowledge transfer facilitates effective adaptation for these low-resource African languages.

***Simba*-SLID.** Following the same ASR setup, we finetune AfriHuBERT, the pretrained model with the broadest African language coverage, using the 215-hour multilingual training split from *SimbaBench*. We validate on the corresponding 21.5-hour development set. This multilingual adaptation supports robust cross-lingual generalization for spoken language identification across diverse African languages.

[7] The CTC layer is a single linear layer placed on top of the pre-trained encoder. For updating all parameters, we perform full parameter fine-tuning, meaning no layers of the base models were frozen during adaptation.

5.3 Evaluation Pipeline

Our evaluation pipeline is designed to ensure consistency across downstream tasks and models, providing a robust framework for analyzing performance under varying resource constraints. As shown in Figure 2, our evaluation pipeline relies on two settings: (i) zero-shot evaluation of baseline models[8], specifically targeting languages not seen during training or not officially supported; and (ii) evaluation of finetuned models to quantify adaptation gains.

For evaluation, we use Word Error Rate (WER) and Character Error Rate (CER) (Woodard and Nelson, 1982; Morris et al., 2004) for ASR, and macro-F_1 (Pedregosa et al., 2011) for SLID. For TTS, we use WER, Mel-Cepstral Distortion (MCD) (Kubichek, 1993), Log F0 Root Mean Square Error (LogF0RMSE) (Lorenzo-Trueba et al., 2018), SpeechTokenDistance (Saeki et al., 2024), Perceptual Evaluation of Speech Quality (PESQ) (Rix et al., 2001), UTMOS (a predicted Mean Opinion Score, MOS) (Reddy et al., 2021), and SpeechBERTScore (Saeki et al., 2024). Detailed information about the experimental setup, hyperparameters, and evaluation metrics is provided in Appendix E.

6 Results

ASR Results. Table 2 presents the performance of baseline systems and our *Simba*-ASR models on *SimbaBench* across 56 language-specific test sets representing 46 languages on the ASR task. Among the 23 test sets for which none of the baseline models officially support, MMS achieves the best performance across all baselines. This trend

[8] These models are already fine-tuned on task-specific data; however, we refer to this as zero-shot since we evaluate them on languages that are unsupported or unseen during training.

is particularly evident for languages such as *Standard Moroccan Tamazight*, *Venda*, *Tswana*, *Swati*, *Sotho*, and *Northern Ndebele*. However, several languages remain challenging for all evaluated models. Specifically, *Susu*, *Tigre*, *Tigrinya*, and *Ga* consistently yield high error rates, revealing substantial gaps in support for certain under-resourced languages. Our finetuned *Simba*-ASR models improve upon every test sets compared to the baseline systems, with *Simba*-S achieving the best overall performance, reaching 41.65 WER and 18.30 CER. These improvements underscore the effectiveness of model adaptation for African languages, with significant improvements for several previously unsupported languages like *Fanti*, *Venda*, *Swati*, and *Bemba*. However, certain languages—including *Western Maninkakan*, *Tigrinya*, *Standard Moroccan Tamazight*, and *Susu*—continue to exhibit high error rates (exceeding 100 WER), indicating that further progress will require additional data and more targeted modeling strategies.

TTS Results. Table 3 presents results on the TTS task across both supported and unsupported languages, across 8 different metrics (More information regarding metrics is provided in Appendix).

MMS-TTS model demonstrates relatively strong performance on its officially supported languages. Hausa stands out with a low Word Error Rate (WER) of 14.09% and a moderate Mel-Cepstral Distortion (MCD) of 8.7. This strong performance is further corroborated by its high scores in human-rated naturalness (3.76 UTMOS) and semantic similarity (0.89 SpeechBERTScore). Ewe also performs well (15.94% WER, 8.87 MCD), though its perceptual quality scores are lower. Performance is competitive for Yoruba (26.99% WER), but drops significantly for Kinyarwanda, which records a high WER of 44.75% and the lowest perceptual quality score in its group (0.68 PESQ), highlighting that synthesis quality can vary considerably even among supported languages. Our finetuned Simba-TTS models, despite limited training data, achieve reasonable results. Nevertheless, intelligibility as measured by WER remains a challenge: 78.31% for Afrikaans, 71.98% for Xhosa, and over 59% for the Twi dialects. Interestingly, we observe improved performance on data derived from BibleTTS (*Lingala*, *Twi Asante*, and *Twi Akuapem*), likely due to the domain's relatively constrained linguistic structure and vocabulary, which appear to support more consistent synthesis.

SLID Results. Table F.2 reports SLID performance across 32 language-dataset pairs using MMS-LID-1024 and *Simba*-SLID. While MMS performs well on high-resource languages, *Simba*-SLID shows notable gains on low-resource languages, addressing key identification gaps.

7 Discussions

Dataset Variation. We find that dataset variations strongly impact performance. On the ASR task, MMS and Seamless models show significantly better performance on *Afrikaans* CV-19 compared to Lwazi and NCHLT datasets. Additionally, both *Zulu* and *Xhosa* consistently achieve better performance on NCHLT datasets than on the Lwazi datasets. This performance gap likely stems from dataset quality differences: NCHLT features broadband speech recordings with over 50 hours per language, while Lwazi contains telephone speech recordings with only 4-10 hours per language, providing more diverse, higher-quality training material in NCHLT. On the SLID task, *Hausa* scores 100% on OlongoAfrica but only 75% on UDHR. This discrepancy likely stems from domain differences: UDHR contains human rights declarations with specialized vocabulary that might complicate language identification, while OlongoAfrica features short stories with more natural language patterns that preserve distinctive linguistic features, making identification easier. Similarly, on the TTS task, test sets drawn from the Bible domain consistently yield lower error rates than those from other domains such as KinyarwandaTTS or SouthAfricaTTS, underscoring the strong influence of domain characteristics on model performance. This reinforces the need for diverse, representative test sets when evaluating multilingual models.

Model Task Coverage. Notably, sheer language coverage does not guarantee uniformly strong ASR accuracy. MMS, which supports the largest number of African languages, attains the best overall average, confirming that extensive pretraining across many languages yields broad, reliable results. Yet this advantage does not extend to every high-resource language: for *Amharic* and *Afrikaans*, Seamless—with far smaller coverage—occasionally surpasses MMS, suggesting that focused training and larger model size can overcome limited coverage when sufficient in-domain data exist. Conversely, Whisper, covering only nine African languages, records the highest error rates

Setting	Language	Test Set	WER (↓)	MCD (↓)	LogF0RMSE (↓)	SpeechTokenDistance (↓)	PESQ (↑)	UTMOS (↑)	SpeechBLEU (↑)	SpeechBERTScore (↑)
MMS-TTS	Ewe (ewe)	bibleTTS	15.94	8.87	0.48	0.57	1.5	3.01	0.51	0.78
	Yoruba (yor)	bibleTTS	26.99	6.72	0.2	0.7	2.61	3.45	0.6	0.89
	Hausa (hau)	bibleTTS	14.09	8.71	0.37	0.52	0.94	3.76	0.39	0.76
	Kinyarwanda (kin)	KinyarwandaTTS	44.75	9.22	0.3	0.57	0.68	3.28	0.4	0.77
	Average		25.44	8.38	0.34	0.59	1.43	3.38	0.48	0.80
Simba-TTS	Xhosa (xho)	SouthAfricaTTS	71.98	7.75	0.29	0.59	0.72	3.1	0.42	0.77
	Lingala (lin)	bibleTTS	35.97	5.12	0.32	0.79	1.51	3.92	0.68	0.89
	Twi Asante (aka)	bibleTTS	59.84	8.3	0.32	0.62	1	3.06	0.48	0.77
	Twi Akuapem (aka)	bibleTTS	59.45	7.19	0.29	0.65	1.05	2.79	0.49	0.81
	Afrikaans (afr)	SouthAfricaTTS	78.31	8.06	0.31	0.59	0.5	3.38	0.41	0.75
	Tswana (tsn)	SouthAfricaTTS	90.3	4.29	0.36	0.6	0.66	2.52	0.4	0.79
	Southern Sotho (sot)	SouthAfricaTTS	91.84	4.28	0.36	0.59	0.74	2.65	0.4	0.78
	Average		69.67	6.43	0.32	0.63	0.88	3.06	0.47	0.79
Overall Average			47.56	7.41	0.33	0.61	1.15	3.22	0.48	0.80

Table 3: Performance of the original MMS-TTS on supported languages and finetuned *Simba*-TTS on unsupported languages. Red underline indicates languages that are not supported by the MMS-TTS model. ↑ indicates higher is better, ↓ indicates lower is better.

overall, and its performance collapses for the many languages it does not officially support, underscoring how lack of task-specific training degrades performance. Overall, wide coverage prevents failure on unsupported languages, whereas fine-grained adaptation determines which system performs best among languages that are already supported.

Relation to Language Family. Our analysis reveals that language family relationships significantly influence model performance patterns across tasks. Within the ***Niger-Congo*** family, closely related *Volta-Congo* languages like *Swahili*, *Zulu*, and *Xhosa* demonstrate similar performances, particularly the *Simba* series models. Low-resource languages benefit substantially from relationships in well-represented families, languages from the *Mande* group achieve reasonable performance despite limited training data, likely due to transfer learning from related ***Niger-Congo*** languages. The effect is especially apparent for ***Afro-Asiatic*** languages; *Amharic* performs exceptionally well with *Simba*-X despite moderate training data, suggesting effective cross-lingual knowledge transfer within its family. These patterns indicate that models leverage shared linguistic features within families, confirming that while comprehensive family representation in training data significantly impacts potential performance, the strength of family representation in training data significantly impacts potential performance, especially for lower-resourced languages of well-represented families.

Trade-off: Intelligibility vs. Voice Quality. For the TTS task, we observe a trade-off between intelligibility and voice quality. Languages such as *Hausa* and *Ewe* show strong performance with low WERs (14–16%), making their generated speech highly understandable. However, their moderate MCD scores (8.7–8.9) indicate acceptable but not perfect voice quality. In contrast, languages such as *Southern Sotho*, *Tswana*, and *Lingala* achieve exceptional voice quality, reflected in excellent MCD and LogF0RMSE scores (4.2–5.1 and 0.32–0.36, respectively). This suggests that their generated audio sounds highly natural and closely matches the target speaker's voice. Nevertheless, this comes at the cost of intelligibility, as indicated by the high WERs (36–92%). Overall, these results suggest that while our finetuning can effectively replicate a speaker's vocal characteristics, achieving accurate and intelligible content synthesis simultaneously remains challenging, particularly with limited in-domain training data.

8 Conclusion

In this work, we present *SimbaBench*, a large-scale benchmark covering 61 African languages across core speech downstream tasks. By curating over 8,600 hours of speech data from diverse domains and language families, we enable comprehensive evaluation of multilingual and Africa-centered speech models. Using *SimbaBench*, we finetune the *Simba* series—task-specific models that enhance performance and mitigate language coverage gaps identified in prior baselines, achieving SoTA results on many low-resource languages. We find that, while broad language coverage provides a useful baseline, our analysis shows that model performance is strongly influenced by domain diversity, data quality, and linguistic relatedness. Our findings underscore the importance of multilingual adaptation and language-family-aware training, highlighting *SimbaBench* as a critical tool for advancing inclusive African speech technologies.

9 Limitations

Our study has a number of limitations that highlight important avenues for future work:

1. **Data Availability and Representation Bias.** *SimbaBench* relies solely on publicly available datasets, which reflect existing structural and historical biases in language technology development. Many Indigenous African languages remain severely underrepresented, limiting the benchmark's ability to capture the full spectrum of Africa's linguistic diversity.

2. **Task Coverage.** Our evaluation is restricted to three core ASR, TTS, and SLID due to data availability. Broader downstream tasks such as speech translation, spoken question answering, or spoken dialogue systems are not yet supported and require further dataset development.

3. **Modeling Scope.** We focus on finetuning existing models rather than proposing new architectural innovations or advanced adaptation methods. While our results demonstrate the benefits of task-specific tuning, we do not explore complementary strategies such as self-supervised pretraining, multitask learning, or data augmentation.

4. **Implementation Constraints.** Despite our advocacy for inclusive data and policy reform, real-world implementation requires sustained institutional commitment. Bridging the gap between research and impact will necessitate long-term investment from governments, academia, and industry partners.

5. **Task Diversity and Generalization.** Although *SimbaBench* spans three speech tasks, it does not yet cover interactive or generative applications such as conversational AI, spoken retrieval, or end-to-end multilingual agents. Extending the benchmark to include such tasks would further promote holistic model evaluation and real-world applicability.

Despite these limitations, our work emphasizes the urgency of addressing speech data disparities and fostering inclusive language technologies across the African continent.

10 Ethical Considerations

We outline several ethical considerations relevant to this work:

1. Our research aims to advance speech technology for African languages by addressing the historical marginalization of many linguistic communities and promoting equitable digital inclusion across the continent.

2. The datasets used in our benchmark are sourced from publicly available repositories. However, their existence reflects broader sociopolitical dynamics, including which languages have received institutional support and technological investment. This highlights the role of policy in shaping digital language presence.

3. Although we do not propose novel model architectures, we fine-tune existing models on *SimbaBench* and release stronger task-specific checkpoints. Our analysis illustrates how unequal data availability—shaped by historical and policy-driven neglect—affects performance, underscoring the need for targeted policy interventions to support multilingual data creation and ethical development.

4. We stress the importance of proper attribution for both datasets and models, as a matter of transparency, accountability, and fair recognition. To this end, we provide a publicly accessible reference list citing all datasets and fine-tuned models used in our benchmark, and encourage researchers and institutions to uphold responsible and inclusive data stewardship.

Acknowledgments

We acknowledge support from Canada Research Chairs (CRC), CLEAR Global for funding from the Gates Foundation, the Natural Sciences and Engineering Research Council of Canada (NSERC; RGPIN-2018-04267), the Social Sciences and Humanities Research Council of Canada (SSHRC; 895-2020-1004; 895-2021-1008), Canadian Foundation for Innovation (CFI; 37771), Digital Research Alliance of Canada,[9] and UBC ARC-Sockeye.[10] We also thank Hellina Hailu Nigatu, Atnafu Lambebo, and Wei-Rui Chen for discussions

[9]https://alliancecan.ca

[10]https://arc.ubc.ca/ubc-arc-sockeye

related to this work. The findings and conclusions contained within this work are those of the authors and do not necessarily reflect positions or policies of any supporters.

References

Gilles Adda, Sebastian Stüker, Martine Adda-Decker, Odette Ambouroue, Laurent Besacier, David Blachon, Helene Bonneau-Maynard, Pierre Godard, Fatima Hamlaoui, Dmitry Idiatov, Guy Noel Kouarata, Lori Lamel, Emmanuel-Moselly Makasso, Annie Rialland, Mark Velde, François Yvon, and Sabine Zerbian. 2016. Breaking the unwritten language barrier: The bulb project. *Procedia Computer Science*, 81:8–14.

Ife Adebara and Muhammad Abdul-Mageed. 2022. Towards afrocentric NLP for African languages: Where we are and where we can go. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3814–3841, Dublin, Ireland. Association for Computational Linguistics.

Ife Adebara, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2024. Cheetah: Natural language generation for 517 african languages. *arXiv preprint arXiv:2401.01053*.

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Inciarte. 2022a. AfroLID: A neural language identification tool for African languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1958–1981, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. 2022b. Serengeti: Massively multilingual language models for africa. *arXiv preprint arXiv:2212.10785*.

Ife Adebara, Hawau Olamide Toyin, Nahom Tesfu Ghebremichael, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2025. Where are we? evaluating llm performance on african languages. *Preprint*, arXiv:2502.19582.

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba O Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, et al. 2024. Irokobench: A new benchmark for african languages in the age of large language models. *arXiv preprint arXiv:2406.03368*.

Jesujoba O Alabi, Xuechen Liu, Dietrich Klakow, and Junichi Yamagishi. 2024. Afrihubert: A self-supervised speech representation model for african languages. *arXiv preprint arXiv:2409.20201*.

Antonios Anastasopoulos, Angela Fan, Dani Haziza, et al. 2023. Seamlessm4t: Massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.04760*.

D. Asamoah Owusu, A. Korsah, B. Quartey, S. Nwolley Jnr., D. Sampah, D. Adjepon-Yamoah, and L. Omane Boateng. 2022. Github - ashesi-org/financial-inclusion-speech-dataset: A speech dataset to support financial inclusion created by ashesi university and nokwary technologies with funding from lacuna fund. https://github.com/Ashesi-Org/Financial-Inclusion-Speech-Dataset.

Arun Babu, Atma Tjandra, Kushal Lakhotia, Apoorv Chauhan, Qiantong Wang, Naman Goyal, Vineel Pratap Jain, Vitaliy Liptchinsky, Ahmed El-Kishky, Juan Pino, Abdelrahman Mohamed, and et al. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1032–1044. Association for Computational Linguistics.

Etienne Barnard, Marelie H. Davel, Charl van Heerden, Febe de Wet, and Jaco Badenhorst. 2014. The nchlt speech corpus of the south african languages. In *Proceedings of the 2014 Spoken Language Technologies for Under-resourced Languages (SLTU) Workshop*, pages 194–200, St. Petersburg, Russia. A comprehensive multilingual speech dataset for the eleven official languages of South Africa.

Tadesse Destaw Belay, Israel Abebe Azime, Ibrahim Said Ahmad, Idris Abdulmumin, Abinew Ali Ayele, Shamsuddeen Hassan Muhammad, and Seid Muhie Yimam. 2025. AfroXLMR-Social: Adapting Pre-trained Language Models for African Languages Social Media Text. *arXiv preprint arXiv:2503.18247*.

Emily M. Bender. 2011. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6.

Laurent Besacier and Elodie Gauthier. 2023. Alffa_public: African languages factored lattices for automatic speech recognition. https://github.com/getalp/ALFFA_PUBLIC.

Alan W Black. 2019. Cmu wilderness multilingual speech dataset. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5971–5975.

Hervé Bredin and Antoine Laurent. 2021. End-to-end speaker segmentation for overlap-aware resegmentation. In *Proc. Interspeech 2021*, Brno, Czech Republic.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2020. pyannote.audio: neural building blocks for speaker diarization. In *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain.

Hervé Bredin. 2023. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. INTERSPEECH 2023*.

Antoine Caubrière and Elodie Gauthier. 2024. Africa-centric self-supervised pre-training for multilingual speech representation in a sub-saharan context. *arXiv preprint arXiv:2404.02000*.

Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Yiwen Guo, and Irwin King. 2024. Recent advances in speech language models: A survey. *arXiv preprint arXiv:2410.03751*.

Ewald Van der westhuizen and Thomas Niesler. 2018. A First South African Corpus of Multilingual Code-switched Soap Opera Speech. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Digital Umuganda. 2023. AfriSpeech Kinyarwanda Male and Female TTS Datasets. https://huggingface.co/datasets/DigitalUmuganda/afrispeak_kinyarwanda_male_tts_dataset.

Moussa Doumbouya, Lisa Einstein, and Chris Piech. 2021. Using radio archives for low-resource speech recognition: Towards an intelligent virtual assistant for illiterate users. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35.

AbdelRahim Elmadany, Ife Adebara, and Muhammad Abdul-Mageed. 2024. Toucan: Many-to-many translation for 150 african language pairs. *arXiv preprint arXiv:2407.04796*.

Elodie Gauthier, Aminata Ndiaye, and Abdoulaye Guissé. 2024. Kallaama: A transcribed speech dataset about agriculture in the three most widely spoken languages in senegal. In *Proceedings of the Fifth Workshop on Resources for African Indigenous Languages (RAIL) @ LREC-COLING 2024*, Lannion, France; Dakar and Thiès, Sénégal.

Alexander Gutkin, Işın Demirşahin, Oddur Kjartansson, Clara Rivera, and Kọ́lá Túbọ̀sún. 2020. Developing an Open-Source Corpus of Yoruba Speech. In *Proceedings of Interspeech 2020*, pages 404–408, Shanghai, China. International Speech and Communication Association (ISCA).

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, et al. 2024. AudioGPT: Understanding and Generating Speech, Music, Sound, and Talking Head. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23802–23804.

Larry M. Hyman. 2003. African languages and phonological theory. *GLOT International*, 7(6):153–163.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 6282. Association for Computational Linguistics.

Robert Kubichek. 1993. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE pacific rim conference on communications computers and signal processing*, volume 1, pages 125–128. IEEE.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Victor Sanh, Thomas Wolf, Lysandre Mouillet, Teven Le Scao, and Alexander M. Rush. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184. Association for Computational Linguistics.

Jaime Lorenzo-Trueba, Junichi Yamagishi, Tomoki Toda, Daisuke Saito, Fernando Villavicencio, Tomi Kinnunen, and Zhenhua Ling. 2018. The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods. *arXiv preprint arXiv:1804.04262*.

Josh Meyer, David Adelani, Edresson Casanova, Alp Öktem, Daniel Whitenack, Julian Weber, Salomon Kabongo Kabenamualu, Elizabeth Salesky, Iroro Orife, Colin Leong, Perez Ogayo, Chris Chinenye Emezue, Jonathan Mukiibi, Salomey Osei, Apelete Agbolo, Victor Akinode, Bernard Opoku, Olanrewaju Samuel, Jesujoba Alabi, and Shamsuddeen Hassan Muhammad. 2022. Bibletts: a large, high-fidelity, multilingual, and uniquely african speech corpus. In *Interspeech*. ISCA.

T. I. Modipa, M. H. Davel, and F. De Wet. 2015. Implications of Sepedi/English Code Switching for ASR Systems. In *Proceedings of the Pattern Recognition Association of South Africa (PRASA)*, pages 112–117.

Andrew Cameron Morris, Viktoria Maier, and Phil D Green. 2004. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Interspeech*, pages 2765–2768.

Mozilla Foundation. 2023. Mozilla common voice: A massively multilingual open dataset for voice technologies. https://commonvoice.mozilla.org.

NaijaVoices. 2024. Naijavoices dataset: A multilingual speech corpus for nigerian languages. https://naijavoices.com/.

Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoit Sagot, Abdelrahman Mohamed, et al. 2023. Generative spoken dialogue language modeling. *Transactions of the Association for Computational Linguistics*, 11:250–266.

Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R Costa-Jussa, Maha Elbayad, Sravya Popuri, Christophe Ropers, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, et al. 2025. Spirit-lm: Interleaved spoken and written language model. *Transactions of the Association for Computational Linguistics*, 13:30–52.

Sewade Ogun, Abraham T Owodunni, Tobi Olatunji, Eniola Alese, Babatunde Oladimeji, Tejumade Afonja, Kayode Olaleye, Naome A Etori, and Tosin Adewumi. 2024. 1000 african voices: Advancing inclusive multi-speaker multi-accent speech synthesis. In *Interspeech 2024*.

Jessica Ojo, Kelechi Ogueji, Pontus Stenetorp, and David Ifeoluwa Adelani. 2023. How good are large language models on african languages? *arXiv e-prints*, pages arXiv–2311.

Akintunde Oladipo, Mofetoluwa Adeyemi, Orevaoghene Ahia, Abraham Toluwalase Owodunni, Odunayo Ogundepo, David Ifeoluwa Adelani, and Jimmy Lin. 2023. Better quality pre-training data and t5 models for African languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 158–168, Singapore. Association for Computational Linguistics.

Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, Sahib Singh, Bonaventure FP Dossou, Joanne Osuchukwu, Salomey Osei, Atnafu Lambebo Tonja, Naome Etori, et al. 2023. AfriSpeech-200: Pan-African Accented Speech Dataset for Clinical and General Domain ASR. *arXiv preprint arXiv:2310.00274*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Alexis Plaquet and Hervé Bredin. 2023. Powerset multi-class cross entropy loss for neural speaker diarization. In *Proc. INTERSPEECH 2023*.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.

Vineel Pratap, Ann Lee, Qiantong Xu, Anuroop Sriram, Tatiana Likhomanenko, Brandon Sottile, et al. 2023. Scaling speech technology to 1,000+ languages. In *Proc. Interspeech*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Chandan KA Reddy, Vishak Gopal, and Ross Cutler. 2021. Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6493–6497. IEEE.

Machel Reid, Junjie Hu, Graham Neubig, and Yutaka Matsuo. 2021. AfroMT: Pretraining strategies and reproducible benchmarks for translation of 8 African languages. *arXiv preprint arXiv:2109.04715*.

Sandy Ritchie, You-Chi Cheng, Mingqing Chen, Rajiv Mathews, Daan van Esch, Bo Li, and Khe Chai Sim. 2022. Large vocabulary speech recognition for languages of africa: multilingual modeling and self-supervised learning. *arXiv preprint arXiv:2208.03067*.

Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. 2001. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, volume 2, pages 749–752. IEEE.

Takaaki Saeki, Soumi Maiti, Shinnosuke Takamichi, Shinji Watanabe, and Hiroshi Saruwatari. 2024. Speechbertscore: Reference-aware automatic evaluation of speech generation leveraging nlp evaluation metrics. *arXiv preprint arXiv:2401.16812*.

Claytone Sikasote and Antonios Anastasopoulos. 2022. BembaSpeech: A Speech Recognition Corpus for the Bemba Language. In *Proceedings of the Language Resources and Evaluation Conference*, pages 7277–7283, Marseille, France. European Language Resources Association.

Claytone Sikasote, Kalinda Siaminwe, Stanly Mwape, Bangiwe Zulu, Mofya Phiri, Martin Phiri, David Zulu, Mayumbo Nyirenda, and Antonios Anastasopoulos. 2023. Zambezi Voice: A Multilingual Speech Corpus for Zambian Languages. In *Proc. INTERSPEECH 2023*, pages 3984–3988.

Kathleen Siminyu, Godson Kalipe, Davor Orlic, Jade Abbott, Vukosi Marivate, Sackey Freshia, Prateek Sibal, Bhanu Neupane, David I Adelani, Amelia Taylor, et al. 2021. Ai4d–african language program. *arXiv preprint arXiv:2104.02516*.

The Brick House Cooperative. 2024. Olongoafrica multilingual anthology. `https://lingua.olongoafrica.com/`. A collection of translated and narrated short stories in various African languages, including Edo, Tamazight, Yoruba, Swahili, Hausa, Tiv, Shona, Ibibio, Igbo, and Nigerian Pidgin.

Hawau Olamide Toyin, Amirbek Djanibekov, Ajinkya Kulkarni, and Hanan Aldarmaki. 2023. ArTST: Arabic text and speech transformer. In *Proceedings of ArabicNLP 2023*, pages 41–51, Singapore (Hybrid). Association for Computational Linguistics.

Universal Declaration of Human Rights Audio. 2025. Universal declaration of human rights audio project. https://udhr.audio/. A project providing audio recordings of the Universal Declaration of Human Rights in multiple languages to promote accessibility and linguistic diversity.

Jörgen Valk and Tanel Alumäe. 2021. Voxlingua107: A dataset for spoken language recognition. In *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*.

Charl Van Heerden, Neil Kleynhans, and Marelie H. Davel. 2016. Improving the lwazi asr baseline. In *Proceedings of Interspeech 2016*.

Daniel van Niekerk, Charl van Heerden, Marelie Davel, Neil Kleynhans, Oddur Kjartansson, Martin Jansche, and Linne Ha. 2017. Rapid development of TTS corpora for four South African languages. In *Proc. Interspeech 2017*, pages 2178–2182, Stockholm, Sweden.

Voice of Africa. 2025. Voice of africa. https://thevoiceofafrica.com/about/. A multilingual platform delivering news and stories from across the African continent.

Jiayi Wang, David Ifeoluwa Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, Sofia Bourhim, Andiswa Bukula, et al. 2023. AfriMTE and AfriCOMET: Enhancing COMET to embrace under-resourced African languages. *arXiv preprint arXiv:2311.09828*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

JP Woodard and JT Nelson. 1982. An information theoretic measure of speech recognition performance. In *Workshop on standardisation for speech I/O technology, Naval Air Development Center, Warminster, PA*.

Marcely Zanon Boito, Vivek Iyer, Nikolaos Lagos, Laurent Besacier, and Ioan Calapodescu. 2024. mhubert-147: A compact multilingual hubert model. In *Interspeech 2024*, pages 3939–3943.

Appendices

The following appendices provide comprehensive supplementary material supporting the main findings of this work. We include detailed descriptions of the datasets used, data preprocessing steps, baseline models, experimental setup, and full evaluation results across tasks and languages. This material is intended to enhance reproducibility, offer deeper insight into model behavior, and serve as a resource for future research in African speech technologies. The appendices are organized as follows:

- §A: Data Collection and Corpus Curation
- §B: Mapping the Data Landscape
- §C: Preprocessing Pipeline
- §D: Baseline Models
- §E: Experimental Setup
- §F: Evaluation Results

Key tables include:

- Table B.1: Total duration of audio (in hours) available per language across multiple datasets.
- Table D.1: Overview of African language coverage across models for pretraining and downstream speech and language tasks.
- Table F.1: Comparison of ASR performance across various African languages using baseline models and our Simba models in both zero-shot and fine-tuned settings.
- Table F.2: Performance of MMS-LID-1024 and Simba-SLID on SimbaBench.

A Data Collection and Corpus Curation

A.1 Automatic Speech Recognition Data

ALFFA PUBLIC Dataset (Besacier and Gauthier, 2023): is a multilingual dataset developed as part of the ALFFA (African Languages in the Field: Speech Fundamentals and Automation) project. It supports ASR systems for under-resourced Sub-Saharan African languages and includes resources for Wolof (5.74 hours), Fongbe (2.89 hours), Amharic (3.12 hours), and Swahili (3.93 hours).

Bemba Speech Dataset (Sikasote and Anastasopoulos, 2022) consists of read speech compiled from various publicly available Bemba sources, including books, show transcripts, and YouTube transcripts. It contains 15,000 utterances totaling 24.5 hours of audio, making it a valuable resource for ASR and linguistic research for the Bemba language.

Mozilla Common Voice (Mozilla Foundation, 2023) is a multilingual dataset designed to improve voice technologies for under-resourced languages. The African language collection includes significant contributions in a variety of languages, with notable amounts of recorded hours in Kinyarwanda (1,354.02 hours), Kabyle (174.66 hours), Ganda (149.43 hours), and Swahili (106.89 hours). Additional contributions include Kalenjin (29.88 hours), Luo (13.63 hours), Hausa (3.77 hours), Taita (4.47 hours), and smaller datasets for languages like Amharic (1.58 hours), Basaa (2.19 hours), and Standard Moroccan Tamazight (1.07 hours). This dataset provides a valuable resource for ASR systems and other linguistic technologies aimed at African languages. More information is available on Common Voice's official page[11].

Financial Inclusion Speech Dataset (Asamoah Owusu et al., 2022) is a multilingual speech dataset developed to support financial inclusion in Ghana. Created by Ashesi University and Nokwary Technologies, the dataset comprises recordings from approximately 200 speakers per language, each recording around 130 sentences. The languages covered include Akuapem Twi (38 hours), Asanti Twi (30 hours), Fanti (39 hours), and Ga (40 hours), totaling approximately 148 hours of speech data.

Kallaama (Gauthier et al., 2024) the Kallaama dataset is a rich resource of transcribed agricultural speech in Senegal's three most widely spoken languages: Wolof, Pulaar, and Sereer. Comprising more than 100 hours of spontaneous audio recordings from farmers, agricultural advisers, and agribusiness managers, the data include radio programs, focus groups, voice messages, and interviews.

Lwazi Speech Corpus (Van Heerden et al., 2016) is a multilingual dataset that includes telephone speech recordings in the 11 official lan-

[11] https://commonvoice.mozilla.org

guages of South Africa. Each language has approximately 200 speakers, each speaker reading an average of 30 prompts, resulting in 4 to 10 hours of audio per language.

NaijaVoices Dataset (NaijaVoices, 2024) is a multilingual speech corpus designed to support ASR and NLP tasks in Nigerian languages. It includes approximately 1,800 hours of speech data and curated text in Yoruba, Igbo, and Hausa, with roughly 600 hours dedicated to each language.

NCHLT Speech Corpus (Barnard et al., 2014) is a multilingual dataset of broadband speech collected from approximately 200 speakers per language in each of the 11 official languages of South Africa: Afrikaans, English, Ndebele, Northern Sotho, Southern Sotho, Swati, Tswana, Tsonga, Venda, Xhosa, and Zulu. Developed under the National Center for Human Language Technology (NCHLT) initiative, the corpus comprises more than 50 hours of orthographically transcribed speech for each language.

Nicolingua - West African Virtual Assistant Speech Recognition Corpus (Doumbouya et al., 2021) is a multilingual dataset comprising 10,083 recorded utterances in four languages: Susu (51 hours), Western Maninkakan (42 hours), Pular (31 hours), and French. Collected from 49 speakers, the corpus is designed to support the development of speech recognition systems for West African languages.

Yoruba Speech Dataset (Gutkin et al., 2020) is a high-quality crowdsourced dataset of Yoruba audio recordings designed for speech processing applications. It includes transcribed WAV files, with separate archives for female and male speakers and the corresponding transcription. It is manually quality-checked and provides valuable resources for developing ASR systems and other linguistic tools for Yoruba.

Zambezi Voice Project (Sikasote et al., 2023) led by the University of Zambia speech and language research group, this ongoing initiative aims to create speech and language resources for Zambia's under-resourced native languages. The labeled dataset comprises more than 36,000 read-speech recordings totaling 79 hours, with contributions from Bemba (26 hours), Nyanja (25 hours), Tonga (22 hours), and Lozi (6 hours). The unlabeled dataset, derived from radio broadcasts, provides 525 hours of audio, including Bemba (162 hours), Tonga (101 hours), Lozi (30 hours), Nyanja (25 hours), and Lunda (39 hours). These resources support the development of ASR and other language technologies for Zambian languages.

A.2 Text-To-Speech Data

BibleTTS (Meyer et al., 2022): BibleTTS is a high-quality multilingual TTS corpus featuring up to 80 hours of studio-quality recordings for each of six Sub-Saharan African languages: Asante Twi, Akuapem Twi, Ewe, Hausa, Lingala, and Yoruba. Derived from the Biblica open.bible project, the dataset includes verse-aligned and filtered speech-text pairs.

High quality TTS data for four South African Languages (van Niekerk et al., 2017): Collected in collaboration between North-West University and Google, this dataset provides over 3 hours of high-quality, multi-speaker transcribed audio recordings for each of the four South African languages: Afrikaans, Sesotho, Setswana, and isiXhosa.

Kinyarwanda TTS (Digital Umuganda, 2023): is a high-quality Text-to-Speech corpus developed and hosted by Digital Umuganda on Hugging Face. The combined dataset totals approximately 14 hours of speech data, covering diverse phonetic contexts and speaking styles.

A.3 Spoken Language Identification Data

NicoLingua - West African Radio Corpus (Doumbouya et al., 2021): This dataset contains 17,090 audio clips, each 30 seconds long, sampled from archives of Guinean radio stations. It spans 10 languages—French, Guerze, Koniaka, Kissi, Kono, Maninka, Mano, Pular, Susu, and Toma—totaling approximately 143.76 hours of audio. The recordings feature a variety of content, including news and radio shows, with rich acoustic diversity such as phone calls, background music, and environmental noise. A validation set of 300 manually tagged clips is included to support evaluation.

VoxLingua107 Dataset (Valk and Alumäe, 2021): is a large-scale multilingual spoken language recognition (SLR) corpus containing over 4,000 hours of YouTube speech data, automatically labeled using language-specific queries. It covers 107 languages and is freely available for research. The dataset includes African languages such as Swahili (57.48h), Somali (92.47h), Shona (27.19h), Amharic (73.36h), Hausa (83.80h), Yoruba (84.66h), Lingala (81.31h), Afrikaans (97.46h), and Malagasy (98.27h). In addition, we specifically selected high-resource non-African languages including Italian (45.91h), Portuguese (58.03h), Spanish (34.95h), Arabic (52.88h), and English (43.84h).

A.4 New Raw Audio Data

OlongoAfrica Multilingual Anthology (The Brick House Cooperative, 2024): is a collection of translated and narrated short stories in 10 African languages, showcasing the linguistic diversity of the continent. The included languages are Edo, Tamazight, Yoruba, Swahili, Hausa, Tiv, Shona, Ibibio, Igbo, and Nigerian Pidgin.

UDHR (Universal Declaration of Human Rights Audio, 2025): The website UDHR.audio hosts raw audio recordings of the Universal Declaration of Human Rights (UDHR) in numerous languages. These recordings capture the text being read aloud by native speakers, Among the languages included, we specifically collected high-quality recordings for Hausa, Tem, Amharic, Wolof, Swahili, and Afrikaans.

VOA (Voice of Africa, 2025)[12]: Voice of Africa includes a collection of news websites delivering updates and stories from across the African continent. This dataset features meticulously collected news videos from the platform in languages such as Tigrinya, North Ndebele, Swahili, Oromo, Kinyarwanda, Somali, Hausa, Amharic, French, Shona, and Lingala, totaling over 1500 hours of speech content.

A.5 Code-Switched Audio Data

CS Soap Opera (der westhuizen and Niesler, 2018): is a multilingual speech dataset compiled from South African soap operas, featuring code-switched speech between English and four Bantu languages: Zulu (5.45h), Xhosa (3.13h), Swana (2.86h), and Southern Sotho (2.83h). It includes multiple forms of code-switching, including between sentences, within sentences, and within individual words—making it a rich resource for studying multilingual ASR in the South African context.

SPCS (Modipa et al., 2015): is a 10.48-hour speech dataset featuring code-switched utterances between Sepedi and English. It was created to support ASR research on multilingual speech involving a minority Bantu language and captures natural switching patterns across diverse speakers and contexts.

[12] https://www.voaafrica.com/

B Mapping the Data Landscape

Table B.1 provides detailed information on the total audio duration (in hours) available for each language across various datasets.

C Preprocessing Pipeline

Audio Standardization. All recordings were resampled to a uniform sampling rate of 16 kHz [13] and converted to single-channel (mono) WAV format. This step ensures compatibility across toolkits and mitigates discrepancies caused by varying source formats and encodings.

Segmentation, Filtering, and Noise Removal. For long-form audio—particularly in unlabeled or newly collected data—we applied silence- and energy-based segmentation to break recordings into utterances. We retained segments with durations between 1 and 20 seconds to avoid instability caused by very short or excessively long samples. To further enhance quality, we removed segments with excessive background noise using energy-based filters. Additionally, we applied voice activity detection (VAD) and speaker diarization using pretrained pipelines from the `pyannote-audio` library, including the voice Activity detection (Bredin et al., 2020; Bredin and Laurent, 2021) and speaker diarization (Plaquet and Bredin, 2023; Bredin, 2023) models.

Metadata Consolidation. All processed datasets were reformatted into a unified JSON-based schema compatible with the Hugging Face `datasets` library (Lhoest et al., 2021) and `fairseq` framework. Each entry includes metadata fields such as audio path, transcription (if available), language ID, dataset origin, and usage split.

D Baseline Models

Whisper-v3 (Radford et al., 2022): Developed by OpenAI, Whisper-v3 is a large-scale encoder-decoder model trained on 680k hours of multilingual and multitask supervised data. We evaluate

[13] A higher sampling rate would be better if we are seeking best settings for the task of TTS.

Language	ISO-3	Hours	Dataset Breakdown (Color-coded)
Afrikaans	afr	255.3	(■, 4.28) (■, 138.73) (■, 3.31) (■, 0.15) (■, 108.39) (■, 0.43)
Akuapim-twi	aka	98.92	(■, 60.57) (■, 38.35)
Asante-twi	aka	31.96	(■, 1.53) (■, 30.43)
Amharic	amh	107.78	(■, 20.76) (■, 0.05) (■, 3.91) (■, 81.47) (■, 1.59)
Basaa	bas	2.19	(■, 2.19)
Bemba	bem	92.81	(■, 26.93) (■, 65.88)
Taita	dav	4.47	(■, 4.47)
Dyula	dyu	0.32	(■, 0.32)
Edo	bin	0.18	(■, 0.18)
Ewe	ewe	77.63	(■, 77.63)
Fanti	fat	39.84	(■, 39.84)
Fon	fon	7.18	(■, 7.18)
Pulaar	fuc	24.8	(■, 24.8)
Pular	fuf	0.52	(■, 0.21) (■, 0.31)
Ga	gaa	40.93	(■, 40.93)
Hausa	hau	908.42	(■, 617.95) (■, 0.14) (■, 0.19) (■, 106.5) (■, 93.3) (■, 86.57) (■, 3.78)
Ibibio	ibb	0.31	(■, 0.31)
Igbo	ibo	634.95	(■, 634.59) (■, 0.33) (■, 0.02)
Kabyle	kab	174.66	(■, 174.66)
Tem	kdh	0.29	(■, 0.29)
Kinyarwanda	kin	1374.35	(■, 14.08) (■, 6.25) (■, 1354.01)
Kalenjin	kln	29.87	(■, 29.87)
Guerze/Kpelle	kpe	0.09	(■, 0.09)
Kisi	kss	0.05	(■, 0.05)
Lingala	lin	201.62	(■, 56.1) (■, 90.26) (■, 55.26)
Lozi	loz	21.64	(■, 6.22) (■, 15.42)
Ganda	lug	149.42	(■, 149.43)
Lunda	lun	20.47	(■, 20.47)
Luo (Kenya and Tanzania)	luo	13.62	(■, 13.62)
Konyanka Maninka	mku	0.12	(■, 0.12)
Malagasy	mlg	109.21	(■, 109.21)
Western Maninkakan	mlq	0.42	(■, 0.42)
Mandinka	mnk	0.63	(■, 0.63)
South Ndebele	nbl	223.88	(■, 4.28) (■, 219.6)
North Ndebele	nde	14.05	(■, 14.05)
Northern Sotho (Sepedi)	nso, eng-nso	188.43	(■, 4.28) (■, 173.66) (■, 0.0) (■, 10.48, CS - English)
Nyanja	nya	36.51	(■, 25.34) (■, 11.17)
Oromo	orm	34.55	(■, 34.55)
Nigerian Pidgin	pcm	0.21	(■, 0.21)
Shona	sna	39.55	(■, 0.3) (■, 8.97) (■, 30.29)
Somali	som	192.12	(■, 89.32) (■, 102.8)
Southern Sotho	sot, eng-sot	184.67	(■, 4.28) (■, 174.34) (■, 3.22) (■, 2.83, CS - English)
Serer	srr	34.38	(■, 34.38)
Susuami	ssu	0.23	(■, 0.23)
Swati	ssw	307.04	(■, 4.28) (■, 302.76)
Susu	sus	0.51	(■, 0.51)
Swahili	swa	689.27	(■, 0.28) (■, 0.19) (■, 506.27) (■, 63.89) (■, 106.89) (■, 11.75)
Tigre	tig	1.04	(■, 1.04)
Tigrinya	tir	39.2	(■, 39.16) (■, 0.04)
Tiv	tiv	0.27	(■, 0.27)
Tonga (Zambia)	toi	85.72	(■, 22.67) (■, 63.06)
Tswana	tsn, eng-tsn	174.7	(■, 4.28) (■, 164.03) (■, 3.52) (■, 0.0) (■, 2.86, CS - English)
Tsonga	tso	145.24	(■, 4.28) (■, 140.96)
Twi	twi	0.21	(■, 0.21)
Central Atlas Tamazight	tzm	0.26	(■, 0.26)
Venda	ven	209.86	(■, 4.28) (■, 205.58)
Wolof	wol	73.65	(■, 18.97) (■, 54.5) (■, 0.18)
Xhosa	xho, eng-xho	225.42	(■, 4.28) (■, 214.89) (■, 3.11) (■, 0.0) (■, 3.13, CS - English)
Yoruba	yor	738.31	(■, 614.98) (■, 0.12) (■, 94.05) (■, 4.03) (■, 25.13)
Standard Moroccan Tamazight	zgh	1.07	(■, 1.07)
Zulu	zul, eng-zul	197.24	(■, 4.28) (■, 187.5) (■, 0.01) (■, 5.45, CS - English)
Multiple*		142.42	(■, 142.42)
English - Accented	eng	200	(■, 200)

Table B.1: Total duration of audio (in hours) available per language across multiple datasets. Color-coded cells indicate the contributing datasets for each language. *The "Multiple" row refers to unlabeled audio data encompassing the following languages: `kpe`, `kss`, `mku`, `mnk`, `fuf`, and `ssu`.

Dataset Legend:

■ afrispeech-200 ■ Alffa_Public ■ BembaSpeech ■ bibleTTS ■ Common Voice 2019 ■ fin_speech ■ Kallaama ■ KinyarwandaTTS ■ Lwazi ■ NaijaVoice ■ NCHTL / AUX ■ Nicolingua-0003 ■ Nicolingua-0004 ■ OlogoAfrica ■ SouthAfricaTTS ■ CS_Soap_Opera ■ SPCS ■ UDHR ■ VOA ■ VoxLingua ■ YorubaVoice ■ ZambeziVoice ■ ZambeziVoice (ULB)

Type	Language	ISO-3	Whisper-v3	M4T-v2	MMS-1B-All	AfriHubert	mHubert	XLS-R
African	Afrikaans	afr	ST	ST	ST LD	PT	PT	PT
	Akuapim-twi	Akuapim-twi	–	–	–	PT	–	–
	Amharic	amh	ST	ST	ST TS LD	PT	PT	PT
	Asante-twi	Asante-twi	–	–	–	PT	–	–
	Basaa	bas	–	–	ST LD	–	–	–
	Bemba	bem	–	–	ST TS LD	PT	–	–
	Central Atlas Tamazight	tzm	–	–	LD	–	–	–
	Dyula	dyu	–	–	ST TS LD	–	–	–
	Edo	bin	–	–	LD	–	–	–
	Ewe	ewe	–	–	ST TS LD	PT	–	–
	Fanti	fat	–	–	–	–	–	–
	Fon	fon	–	–	ST TS LD	–	–	–
	Ga	gaa	–	–	LD	–	–	–
	Ganda	lug	–	ST	ST TS LD	PT	PT	PT
	Guerze/Kpelle	kpe	–	–	–	–	–	–
	Hausa	hau	ST	–	ST TS LD	PT	PT	PT
	Ibibio	ibb	–	–	LD	–	–	–
	Igbo	ibo	–	ST	ST LD	PT	PT	–
	Kabyle	kab	–	–	ST TS LD	–	PT	PT
	Kalenjin	kln	–	–	–	–	–	–
	Kinyarwanda	kin	–	–	ST TS LD	PT	PT	PT
	Kisi	kss	–	–	ST TS LD	PT	–	–
	Konyanka Maninka	mku	–	–	LD	PT	–	–
	Lingala	lin	ST	–	ST LD	PT	PT	–
	Lozi	loz	–	–	LD	PT	–	–
	Lunda	lun	–	–	LD	PT	–	–
	Luo (Kenya and Tanzania)	luo	–	ST	ST LD	–	–	–
	Malagasy	mlg	ST	–	ST TS LD	PT	–	PT
	Mandinka	mnk	–	–	ST TS LD	PT	–	–
	Nigerian Pidgin	pcm	–	–	ST TS LD	–	–	–
	North Ndebele	nde	–	–	LD	–	–	–
	Northern Sotho (Sepedi)	nso	–	–	ST LD	PT	–	–
	Nyanja	nya	–	ST	ST TS LD	PT	–	–
	Oromo	orm	–	–	ST TS LD	–	–	–
	Pulaar	fuc	–	–	–	–	–	–
	Pular	fuf	–	–	–	PT	–	–
	Serer	srr	–	–	LD	PT	–	–
	Shona	sna	ST	ST	ST TS LD	PT	PT	PT
	Somali	som	ST	ST	ST TS LD	PT	PT	PT
	South Ndebele	nbl	–	–	LD	PT	–	–
	Southern Sotho	sot	–	–	LD	PT	PT	–
	Standard Moroccan Tamazight	zgh	–	–	–	–	–	–
	Susu	sus	–	–	ST TS LD	PT	–	–
	Susuami	ssu	–	–	–	–	–	–
	Swahili	swa, swh	ST	–	ST TS LD	PT	PT	PT
	Swati	ssw	–	–	LD	PT	–	–
	Taita	dav	–	–	–	–	–	–
	Tem	kdh	–	–	ST TS LD	–	–	–
	Tigre	tig	–	–	LD	–	PT	–
	Tigrinya	tir	–	–	ST TS LD	–	–	–
	Tiv	tiv	–	–	LD	–	–	–
	Tonga (Zambia)	toi	–	–	LD	PT	–	–
	Tsonga	tso	–	–	ST TS LD	PT	–	–
	Tswana	tsn	–	–	LD	PT	PT	–
	Twi	twi	–	–	–	–	–	–
	Venda	ven	–	–	LD	PT	–	–
	Western Maninkakan	mlq	–	–	LD	–	–	–
	Wolof	wol	–	–	ST LD	PT	–	–
	Xhosa	xho	–	ST	ST LD	PT	PT	–
	Yoruba	yor	–	ST	ST TS LD	PT	PT	PT
	Zulu	zul	ST	ST	ST LD	PT	–	PT
Code-switched	English - Southern Sotho	eng-sot	–	–	–	–	–	–
	English - Tswana	eng-tsn	–	–	–	–	–	–
	English - Xhosa	eng-xho	–	–	–	–	–	–
	English - Zulu	eng-zul	–	–	–	–	–	–
	Northern Sotho - English	nso-eng	–	–	–	–	–	–
Non-African	English - Accented	eng	ST	ST	ST TS LD	PT	PT	PT

Table D.1: Overview of African language coverage across models for pretraining and downstream speech and language tasks. **Abbreviations**: ST Speech-to-Text (ASR), TS Text-to-Speech, LD Spoken Language Identification, and PT Pretraining.

two variants: `whisper-large-v3`, which offers high accuracy for multilingual ASR tasks, and `whisper-large-v3-turbo`, which provides faster inference with a slight trade-off in accuracy.

SeamlessM4T-v2 Large (Anastasopoulos et al., 2023): A unified model by Meta AI supporting speech-to-text, speech-to-speech, and text-to-text translation across over 100 languages. It is particularly designed for low-latency and zero-shot multilingual translation.

MMS-1b-All (Pratap et al., 2023): Part of Meta's Massively Multilingual Speech (MMS) project, this model is trained on over 1,100 languages with 1B parameters. It supports ASR and SLID tasks and represents the largest multilingual speech pretraining effort to date. Pratap et al. also trained the VITs architecture for TTS on a number of languages, including 3 of the African languages covered in our work.

AfriHUBERT (Alabi et al., 2024): A HuBERT-based self-supervised model trained exclusively on African speech data. It focuses on improving representation learning for low-resource African languages and is optimized for ASR and feature extraction tasks.

Wav2Vec2-XLS-R (Babu et al., 2021): A family of cross-lingual speech representation models developed by Facebook AI, trained using the wav2vec2 framework on a multilingual dataset spanning 128 languages. We evaluate two variants: `facebook/wav2vec2-xls-r-300m` and `facebook/wav2vec2-xls-r-1b`, which differ in parameter count and pretraining scale. These models are widely used for fine-tuning on low-resource ASR tasks due to their strong generalization across languages.

Table D.1 presents a detailed overview of African language support across models for pretraining and various downstream tasks in speech and language processing.

E Experimental Setup

For the *Simba* series of models, we select the best checkpoint for each model based on development set performance at the end of each epoch. These selected checkpoints are then used for final evaluation, during which we compute task-specific metrics and report the results accordingly.

Hyperparameters. All ASR and SLID models are fine-tuned using the Adam optimizer with a cosine learning rate of 5×10^{-5} over 30 epochs. We use the HuggingFace Transformers (Wolf et al., 2020) for training and evaluation.[14] For TTS models, we adopt the finetuning procedure outlined in the Vits repository and follow the default hyperparameter configuration provided in the repository[15].

Evaluation Metrics. For ASR, we evaluate using Word Error Rate (WER) (Woodard and Nelson, 1982; Morris et al., 2004) and Character Error Rate (CER) (Morris et al., 2004).

For SLID, we use macro-F_1 (Pedregosa et al., 2011) to address class imbalance and ensure balanced performance assessment across languages. For TTS, we assess synthesized speech intelligibility with the best available ASR model for each language, reporting both WER and CER as objective measures following (Toyin et al., 2023). Word Error Rate (WER) (Woodard and Nelson, 1982; Morris et al., 2004), measures the accuracy of the synthesized speech by comparing the transcribed output to the original text, where a lower WER indicates fewer errors (insertions, deletions, and substitutions) and thus higher intelligibility. **Mel-Cepstral Distortion (MCD)**(Kubichek, 1993) serves as an objective measure of the difference between the spectral features of the synthesized speech and natural speech, with a lower MCD suggesting that the synthesized speech is acoustically more similar to human speech. **Log F0 Root Mean Square Error (LogF0RMSE)** (Lorenzo-Trueba et al., 2018) evaluates the accuracy of the synthesized speech's pitch (fundamental frequency) compared to a reference, where a lower value indicates more natural and accurate intonation. **SpeechTokenDistance** (Saeki et al., 2024) calculates the distance between sequences of discrete speech tokens from the generated and reference speech, with a smaller distance implying a closer match in the fundamental units of speech. **Perceptual Evaluation of Speech Quality (PESQ)** (Rix et al., 2001) is a standardized algorithm for objectively measuring the perceptual quality of speech, where a higher PESQ score indicates higher perceived quality, often used in telecommunications. **UTMOS is a predicted Mean Opinion Score (MOS)** (Reddy et al., 2021) generated by a machine learning model that aims to replicate human-rated scores for speech

[14] https://github.com/huggingface/transformers
[15] https://github.com/ylacombe/finetune-hf-vits

naturalness, with a higher UTMOS score suggesting a more natural-sounding voice. SpeechBLEU, inspired by the BLEU score in machine translation, measures the similarity of the generated speech to a reference at the level of n-grams of discrete speech tokens, where a higher score indicates better fluency and similarity to the reference. Finally, **SpeechBERTScore** (Saeki et al., 2024) leverages deep learning models (BERT) to compare the semantic similarity between the generated and reference speech, with a higher score suggesting that the meaning and context are well-preserved in the synthesized audio.

F Evaluation Results

Table F.1 shows the detailed results of all models across all AST test sets. Table F.2 presents the results for the spoken language identification task.

Language	Test Set	MMS	Seamless	Whisper	WhisperT	*Simba* Series (Ours)				
						Simba-H	*Simba*-M	*Simba*-S	*Simba*-X	*Simba*-W
Akuapim-twi (aka)	FS	85.82/40.14	219.67/190.49	1181.0/1131.23	499.51/547.24	26.83/10.13	17.6/8.13	13.29/8.45	23.74/10.35	29.1/19.1
Asante-twi (aka)	FS	83.6/32.35	230.88/196.71	665.34/574.27	245.5/222.37	26.78/7.36	13.87/5.38	7.06/2.62	19.93/7.06	15.63/7.98
Afrikaans (afr)	Lwazi	92.06/37.59	37.91/16.47	66.05/34.32	73.17/39.05	62.81/17.9	36.29/9.86	15.62/4.99	102.96/53.45	29.22/11.0
Afrikaans (afr)	NCHTL	118.72/31.86	27.96/4.63	77.61/24.22	67.61/15.2	53.57/8.16	25.55/3.4	12.39/2.01	109.93/36.25	20.82/3.81
Afrikaans (afr)	CV-19	26.29/6.7	19.52/9.18	35.85/9.9	46.38/17.11	64.15/19.97	35.36/13.19	16.97/7.47	93.32/46.55	27.87/11.27
Amharic (amh)	CV-19	51.93/21.81	87.58/22.25	432.1/294.11	245.47/236.28	86.93/42.59	58.26/25.39	42.14/16.94	105.96/119.54	106.34/65.09
Basaa (bas)	CV-19	34.4/9.6	147.17/109.79	554.16/475.04	169.5/123.55	61.08/20.41	36.51/10.27	65.17/24.97	84.09/30.86	76.39/31.3
Bemba (bem)	BS	47.73/7.95	187.48/106.1	921.43/515.91	136.53/70.37	51.9/9.28	44.06/7.1	38.99/7.59	83.32/20.12	50.84/10.51
Taita (dav)	CV-19	82.47/25.19	170.25/104.12	662.71/401.46	151.56/86.05	67.34/20.59	58.49/16.99	44.79/15.29	82.66/27.59	105.83/60.98
Dyula (dyu)	CV-19	65.61/16.14	152.07/104.41	424.53/344.84	107.85/43.71	77.98/23.26	67.99/21.53	78.07/23.11	85.58/26.57	87.02/26.42
Fanti (fat)	FS	115.34/61.98	244.53/209.09	1188.25/1082.92	497.67/581.04	23.38/7.27	19.97/6.99	8.58/4.96	27.89/9.94	23.06/15.66
Fon (fon)	Alffa	87.83/31.92	132.67/115.18	488.58/467.05	159.01/134.6	33.29/9.49	44.51/12.52	43.75/14.77	53.81/17.4	45.54/16.72
Pulaar (fuc)	Kallaama	103.25/68.15	200.49/144.98	904.99/743.08	321.82/280.08	91.74/56.75	87.29/54.54	69.39/43.09	96.67/67.88	107.04/73.33
Pular (fuf)	NL4-WA	106.98/51.34	244.57/177.0	789.15/740.44	553.1/435.86	106.2/50.68	101.55/44.86	98.06/54.05	96.9/53.01	136.05/75.85
Ga (gaa)	FS	139.31/55.26	322.68/230.61	1362.33/1043.95	482.59/412.01	41.56/11.51	20.35/7.35	9.67/6.38	32.97/10.37	22.21/11.99
Hausa (hau)	CV-19	27.63/5.97	135.46/91.76	110.59/56.15	130.84/72.4	55.68/15.58	29.42/6.58	64.19/23.1	92.06/34.61	90.54/47.26
Igbo (ibo)	CV-19	70.82/18.08	61.66/18.02	111.67/62.92	321.93/175.32	89.02/35.57	83.35/26.58	77.73/34.93	98.33/42.57	95.27/54.04
Kabyle (kab)	CV-19	49.49/14.33	149.33/101.52	508.87/412.14	153.96/123.25	79.21/25.64	62.81/16.99	58.78/20.46	93.46/44.89	67.02/29.09
Kinyarwanda (kin)	CV-19	34.22/9.42	167.9/93.77	820.55/473.72	245.92/141.28	55.33/15.82	38.59/10.23	54.22/18.14	91.2/33.29	72.8/24.77
Kalenjin (kln)	CV-19	99.97/34.64	178.09/95.33	773.64/453.35	178.81/107.16	80.44/21.49	73.43/18.86	70.37/18.19	85.26/25.65	75.93/21.24
Lozi (loz)	Z.Voice	87.37/32.2	124.72/95.9	657.46/508.83	109.26/55.17	61.27/23.84	63.58/23.66	57.34/22.92	87.28/32.05	64.39/24.12
Ganda (lug)	CV-19	26.21/5.35	17.69/4.27	866.81/468.58	168.18/77.31	64.15/13.64	35.24/6.37	23.11/5.65	88.66/25.19	55.92/13.83
Luo (luo)	CV-19	111.02/76.27	111.43/53.86	478.84/332.08	115.16/53.15	56.86/13.6	42.4/9.05	38.79/10.29	67.28/16.55	52.18/13.27
W. Maninkakan (mlq)	NL4-WA	113.02/59.21	228.93/171.01	1232.92/1237.86	306.11/217.65	110.62/48.96	98.97/40.52	115.82/51.12	96.65/47.21	176.76/113.79
S. Ndebele (nbl)	Lwazi	74.29/31.76	139.42/91.29	349.57/199.2	100.43/56.06	62.13/18.33	38.44/10.89	19.02/7.4	103.06/52.98	29.58/11.33
S. Ndebele (nbl)	NCHTL	58.61/10.53	238.25/104.08	1368.24/566.76	198.86/86.4	31.95/5.57	33.13/5.45	25.51/4.99	66.75/11.16	36.32/6.14
Northern Sotho (nso)	Lwazi	84.64/32.4	147.45/94.85	251.7/175.66	105.29/71.63	67.06/19.27	43.43/11.37	21.27/7.84	104.06/54.47	33.07/10.22
N. Sotho (nso)	NCHTL	42.69/11.39	154.46/120.61	611.95/512.8	158.31/140.4	20.72/5.21	21.49/5.09	16.39/4.42	47.05/13.71	22.45/6.44
Nyanja (nya)	Z.Voice	99.85/82.25	25.34/7.0	744.72/392.55	92.22/23.12	50.61/10.99	46.8/9.78	22.38/5.99	76.22/18.17	41.61/8.94
S. Sotho (sot)	Lwazi	70.04/29.11	132.03/86.73	248.55/193.15	110.71/52.4	61.59/17.94	38.2/10.41	18.63/7.24	102.48/54.0	31.81/11.55
S. Sotho (sot)	NCHTL	79.97/27.48	154.26/111.44	743.88/591.26	145.42/113.09	23.94/6.31	26.84/6.87	18.15/5.58	44.74/12.54	24.47/7.3
Serer (srr)	Kallaama	105.41/69.85	255.33/233.38	1046.88/977.99	479.84/571.07	95.21/55.41	94.26/56.94	88.39/62.34	96.22/68.31	125.44/113.36
Swati (ssw)	Lwazi	73.08/29.37	139.27/88.48	338.16/309.79	113.7/61.78	64.93/18.42	39.59/10.4	17.94/6.57	101.49/54.47	30.79/11.01
Swati (ssw)	NCHTL	65.0/10.76	247.32/106.42	1345.67/539.36	221.86/77.57	22.88/3.15	29.39/4.14	20.6/3.28	62.45/9.55	34.35/5.07
Susu (sus)	NL4-WA	150.79/123.0	264.17/177.19	665.0/471.49	491.35/496.48	120.4/48.53	107.5/36.83	126.55/51.74	108.81/44.54	215.16/121.32
Swahili (swa)	CV-19	25.65/7.46	15.86/6.16	81.89/38.84	95.0/42.51	42.46/11.6	24.77/7.23	16.52/6.15	68.07/19.14	34.7/11.87
Swahili (swh)	Alffa	40.8/12.37	25.0/10.47	63.9/23.29	63.55/24.44	43.87/11.36	29.29/7.87	16.61/5.64	71.51/22.25	25.84/8.22
Tigre (tig)	CV-19	115.07/120.83	213.66/207.69	690.76/652.88	143.04/179.65	71.94/30.13	59.25/21.02	57.74/26.16	102.46/90.21	87.39/67.43
Tigrinya (tir)	CV-19	117.74/111.01	189.88/180.49	165.36/148.48	135.48/158.64	90.24/47.95	92.5/65.98	75.24/50.84	100.71/91.24	122.5/115.54
Tonga (Zambia) (toi)	Z.Voice	71.71/14.91	188.65/92.13	1175.58/550.42	127.14/41.37	63.02/10.74	42.25/6.82	51.31/8.01	85.57/22.14	57.49/10.66
Tswana (tsn)	Lwazi	72.4/31.33	140.76/93.65	231.9/159.83	119.46/84.15	62.14/17.09	37.45/10.51	18.2/6.64	102.84/53.11	28.44/10.2
Tswana (tsn)	NCHTL	63.26/18.88	165.25/109.5	795.5/551.5	161.95/135.85	18.9/4.26	22.38/4.88	12.95/3.46	44.1/10.84	18.86/4.87
Tsonga (tso)	Lwazi	80.41/33.76	142.02/91.96	264.92/172.82	91.92/55.2	62.69/18.33	38.48/9.98	17.0/6.05	102.67/53.21	34.21/20.76
Tsonga (tso)	NCHTL	61.74/10.46	163.28/107.2	1105.49/748.39	148.77/102.28	22.5/4.0	25.87/4.41	17.77/3.65	55.94/11.75	27.45/6.12
Twi (twi)	CV-19	94.32/42.27	128.78/96.52	599.6/403.21	105.51/52.47	74.97/26.69	81.06/30.01	62.68/15.21	84.81/31.96	91.34/28.7
Venda (ven)	Lwazi	71.16/29.89	140.9/95.47	265.76/220.81	129.56/155.71	62.66/19.31	38.71/11.41	19.13/6.96	102.19/52.9	30.95/11.58
Venda (ven)	NCHTL	85.98/27.41	159.21/112.41	653.82/456.88	122.93/79.61	28.28/6.12	33.11/6.99	27.37/6.89	68.34/20.29	32.21/7.87
Wolof (wol)	Alffa	43.57/10.44	128.76/92.3	446.07/348.0	202.6/143.81	59.7/15.41	34.75/8.03	40.65/13.28	89.3/30.42	34.42/9.82
Wolof (wol)	Kallaama	101.14/81.39	1050.1/1020.36	1050.1/1020.36	374.1/388.4	105.0/75.1	100.44/77.09	100.20/75.38	102.95/82.07	143.16/131.49
Xhosa (xho)	Lwazi	73.89/33.13	140.48/90.83	286.14/227.12	148.78/78.46	67.97/19.69	43.26/11.49	22.1/7.83	101.99/53.45	46.67/38.59
Xhosa (xho)	NCHTL	35.24/5.76	246.81/117.89	1405.1/615.03	217.5/87.08	34.43/5.58	32.33/5.09	28.66/5.28	68.16/11.29	40.82/7.08
Yoruba (yor)	Y.Voice	50.12/18.29	23.47/11.8	639.75/503.98	105.84/70.0	40.59/13.49	41.21/12.91	20.12/11.72	98.51/55.74	52.01/25.45
S. M. Tamazight (zgh)	CV-19	107.34/98.24	150.25/123.89	371.93/326.05	129.36/125.21	102.04/86.11	90.85/72.04	111.43/98.69	101.33/91.43	108.25/94.8
Zulu (zul)	Lwazi	70.12/32.66	107.96/84.77	164.54/106.64	78.11/43.35	62.92/17.57	38.58/10.88	108.53/103.61	101.93/52.87	27.63/10.87
Zulu (zul)	NCHTL	31.31/5.12	74.28/20.56	648.45/244.13	379.87/134.73	30.55/4.69	26.36/3.96	23.87/4.47	60.96/8.79	33.92/5.71
Overall Average		75.9/35.26	146.69/98.92	611.91/437.98	196.7/149.79	59.9/21.46	48.11/17.41	41.65/18.3	82.64/39.31	60.56/31.16

Table F.1: Comparison of ASR performance across various African languages using baseline models and our's *Simba* models in both zero-shot and fine-tuned settings. The evaluation metrics are reported as `WER/CER`. Red Underline indicates that the model does not support the corresponding language. Green indicates the best-performing model for each language/test set. **Abbreviations**: `FS` – Financial Speech, `BS` – Bemba Speech, `CV-19` – Common Voice 2019, `NL4-WA` – Nicolingua-0004-West Africa, `Z.Voice` – Zambezi Voice, `Y.Voice` – Yoruba Voice, `S.M.` – Standard Moroccan, `N.` – Northern, `S.` – South/Southern, `W.` – Westren.

Language	**Test Set**	**MMS-LID-1024**	***Simba*-SLID**
Edo (bin)	OlogoAfrica	6.25	80.12
Afrikaans (afr)	UDHR	88.89	88.89
Amharic (amh)	UDHR	100.00	100.00
Amharic (amh)	VoxLingua	98.50	89.39
Bemba (bem)	ZambeziVoice	26.67	53.15
Hausa (hau)	OlogoAfrica	100.00	100.00
Hausa (hau)	UDHR	75.00	75.00
Hausa (hau)	VoxLingua	97.99	94.18
Ibibio (ibb)	OlogoAfrica	14.29	25.98
Igbo (ibo)	OlogoAfrica	65.79	75.26
Tem (kdh)	UDHR	45.45	55.23
Kinyarwanda (kin)	VOA	20.50	21.92
Lingala (lin)	VoxLingua	96.29	15.86
Lozi (loz)	ZambeziVoice	1.36	5.30
Lunda (lun)	ZambeziVoice	23.60	30.12
Malagasy (mlg)	VoxLingua	98.55	70.12
Nyanja (nya)	ZambeziVoice	22.64	15.59
Nigerian Pidgin (pcm)	OlogoAfrica	73.68	74.32
Shona (sna)	OlogoAfrica	90.91	92.34
Shona (sna)	VoxLingua	86.61	88.23
Somali (som)	VoxLingua	97.96	95.54
Swahili (swa, swh)	OlogoAfrica	99.03	94.14
Swahili (swa, swh)	UDHR	99.60	94.29
Swahili (swa, swh)	VoxLingua	99.96	94.29
Tiv (tiv)	OlogoAfrica	66.67	69.93
Tonga (Zambia) (toi)	ZambeziVoice	31.48	56.47
Central Atlas Tamazight (tzm)	OlogoAfrica	27.27	40.76
Wolof (wol)	UDHR	83.33	83.33
Yoruba (yor)	OlogoAfrica	100.00	100.00
Yoruba (yor)	VoxLingua	96.27	95.87
Overall Average		69.44	70.82

Table F.2: Performance of the MMS-LID-1024 on *SimbaBench* and *Simba*-SLID. Green indicates the best-performing model for each language/test set. The evaluation metrics are reported as $F_1 - macro$.