African Voices Nigeria: 2500 hours of ethically sourced speech data for four Nigerian Languages

Ife Adebara[ξ,Ω,λ] **Oluwaseun Nifemi**[ξ,Ω] **Olubayo Adekanmbi** [ξ,Ω] **Rashidat Damilola Sikiru**[ξ,Ω]
Ololade Anjuwon[ξ,Ω] **Olanrewaju Israel Lawal**[ξ,Ω] **Anthony Soronnadi** [ξ,Ω]
John Emeka Eze [ξ] **Ewezu Ngim Ngim** [ξ]
[ξ]Data Science Nigeria [Ω]EqualyzAI [λ] University of Alberta
Contact: {ife,oluwaseun,olubayo}@equalyz.ai

Abstract

African languages remain severely underrepresented in large-scale speech resources, particularly for spontaneous, naturally occurring speech that reflects real-world linguistic use. We present African Voices, a large-scale, ethically governed speech dataset covering four Nigerian languages, [2,500] hours of audio, and 2,865 speakers, with a focus on spontaneous and scripted speech across diverse sociolinguistic contexts. Unlike existing resources that primarily rely on read or scripted speech, African Voices captures natural variation in accent, dialect, register, and code-switching, accompanied by rich demographic and contextual metadata. We describe the data collection methodology, transcription and a principled governance framework designed to support responsible use of speech data in low-resource settings. We further provide baseline automatic speech recognition results across languages. African Voices enables research on robust and fair ASR and serves as a foundational resource for advancing NLP research in African languages.

1 Introduction

Despite recent advances in multilingual speech and language models (Ardila et al., 2020a; Baevski et al., 2020; Pratap et al., 2024; team et al., 2025), African languages remain systematically underrepresented in the datasets used to train and evaluate speech technologies. Existing large-scale speech resources often prioritize read or scripted speech, limited demographic annotation, and weak governance mechanisms, resulting in models that perform poorly in real-world African contexts. Spontaneous speech—characterized by natural prosody, disfluencies, dialectal variation, and code-switching—poses distinct challenges for automatic speech recognition (ASR) systems, yet remains largely absent from publicly available African language datasets. This gap limits both model robustness and our ability to study fairness, bias, and generalization across sociolinguistic groups (Benzeghiba et al., 2007; Besacier et al., 2015; Nakamura et al., 2008; Koenecke et al., 2020).

To address these limitations, we introduce African Voices, a large-scale speech dataset designed to reflect the linguistic, demographic, and contextual diversity of African language use. It consists of 2500 hours of speech data collected from 2,856 diverse speakers across four languages - Hausa, Igbo, Nigerian Pidgin, and Yoruba. African Voices prioritizes spontaneous speech collected from speakers across age groups, genders, regions, and domains, and is accompanied by fine-grained metadata to support sociolinguistic and fairness-aware analysis. By grounding speech data in real-world linguistic contexts, African Voices enables principled evaluation of ASR systems beyond read speech benchmarks, supports analysis of demographic and dialectal performance disparities, and facilitates research on explainable and robust speech technologies for low-resource languages.

In addition to its technological objectives, the project foregrounds ethical and participatory data practices. All speech data is collected with informed consent, and transcriptions are produced by native and highly proficient speakers, ensuring both linguistic accuracy and cultural sensitivity. By enabling the development of robust ASR and TTS models for African languages, the African Voices Project contributes to the broader goals of linguistic inclusion, digital equity, and sustainable language technology development in underrepresented regions. Our contributions are as follows: **(1.)** We release African Voices[1], a large-scale, multi-language African speech dataset focused on spontaneous and scripted speech. **(2.)** We provide rich demographic and contextual metadata enabling fairness, bias, and sociolinguistic analyses. **(3.)** We

[1]Available https://www.africanvoices.io/

describe a human-in-the-loop anonymization and governance framework tailored to speech data in African contexts. **(4.)**We present baseline ASR results across languages.

This paper is organized as follows. Section 2 surveys related literature on African language speech resources. Section 3 provides linguistic background on Hausa, Igbo, Naija Pidgin, and Yoruba. Section 4 describes our data collection methodology and ethical governance framework. Section 5 presents baseline ASR experiments and error analyses across languages and demographic groups. Section 6 concludes and discusses future research directions.

2 Literature Review

The landscape of automatic speech recognition (ASR) and text-to-speech (TTS) technologies has progressed rapidly in recent years, driven by advances in self-supervised learning and large-scale multilingual modeling. Despite these developments, African languages remain critically underrepresented in both training data and evaluation benchmarks. This exclusion stems from persistent challenges related to data availability, dataset quality, model design assumptions, and evaluation practices that historically privilege high-resource languages. In this section, we synthesize prior work on multilingual speech technology, African-language speech resources, and dataset creation methodologies to contextualize the contribution of the African Voices project.

2.1 African Speech Corpora and Dataset Development

Efforts to develop speech corpora for African languages have accelerated in recent years, though substantial gaps in scale, diversity, and coverage remain. Alfa project (Tachbelie et al., 2014; Gelas et al., 2012; Gauthier et al., 2016) is an early attempt for ASR data collection in Africa. It covers Amharic, Hausa, Swahili, and Wolof. One of the most comprehensive initiatives to date is NaijaVoices (Emezue et al., 2025), which provides 1,867 hours of speech data across Hausa, Igbo, and Yoruba from more than 5,000 speakers. Fine-tuning Whisper models on this corpus yields up to a 75% relative reduction in word error rate (WER), highlighting the impact of high-quality, language-specific data. For African-accented English, AfriSpeech-200 (Olatunji et al., 2023) offers 200.7 hours of speech spanning 120 accents. Evaluation on this dataset reveals marked disparities in ASR performance, with commercial systems exhibiting WERs between 10–25% on African accents compared to sub-10% on native varieties. Similarly, the ÌròyìnSpeech corpus (Ogunremi et al., 2023) contributes 42 hours of multi-domain Yoruba speech, establishing a baseline WER of 23.8% and demonstrating that high-quality TTS systems can be developed with relatively modest but carefully curated data.

Additional single-language resources include the Sagalee Oromo corpus (Abu et al., 2025), comprising approximately 100 hours of speech and achieving a 10.8% WER with Whisper-based models, and the EthioSpeech Corpora, which provide 391 hours across six Ethiopian languages. BibleTTS (Meyer et al., 2022) offers up to 86 hours per language for ten African languages, but is constrained by single-speaker recordings and domain-specific (religious) content, limiting its applicability for general-purpose speech modeling. While these datasets have significantly advanced African-language speech technology, they are often limited to specific domains, scripted or read speech, or narrow speaker distributions, restricting their ability to capture real-world linguistic variability.

2.2 Data Quality and Methodological Challenges

Beyond scale, data quality and methodological rigor remain critical challenges in African speech datasets. Comprehensive audits by Lau et al. (2025) of widely used benchmarks—including Common Voice 17, FLEURS, and VoxPopuli—identify systematic labeling errors and inconsistencies that disproportionately affect low-resource languages. Although Common Voice (Ardila et al., 2020b) covers a broad range of African languages, data quality, demographic balance, and recording conditions vary substantially across languages and locales. The FLEURS benchmark (Conneau et al., 2022), while valuable for multilingual evaluation, provides only approximately 12 hours of speech per language, rendering it insufficient for robust model training or detailed sociolinguistic analysis. Moreover, many existing datasets lack fine-grained demographic and contextual metadata, limiting their utility for studying bias, fairness, and performance variation across speaker groups. These limitations underscore the need for datasets that combine scale, quality control, sociolinguistic diversity, and transpar-

Dataset	Hours	Lang.	Speakers	Diversity
African voices	2500	4	2500+	Multi-domain, multi-style
NaijaVoices (Emezue et al., 2025)	1 867	3	5 000+	Culturally rich prompts
AfriSpeech-200 (Olatunji et al., 2023)	200.7	1	2 463	Clinical + general
BibleTTS (Meyer et al., 2022)	86/lang	10	10	Studio, religious text
ÌròyìnSpeech (Ogunremi et al., 2023)	42	1	80+	News, creative
Sagalee (Abu et al., 2025)	100	1	–	Read, monolingual
AfroTTS (Ogun et al., 2024)	136	1	747	Multi-accent TTS
FLEURS (Conneau et al., 2022)	∼12/ lang	102	–	Parallel eval

Table 1: Comparative Overview of African-Language Speech Datasets, Highlighting the African Voices Corpus

ent documentation—particularly for spontaneous speech, which remains underrepresented despite its importance for real-world ASR robustness.

2.3 Spontaneous Speech and Sociolinguistic Variation

Spontaneous speech poses well-documented challenges for ASR systems due to disfluencies, informal registers, prosodic variability, and code-switching (Godfrey et al., 1992; Adda-Decker et al., 2000). Models trained predominantly on read or scripted speech frequently exhibit degraded performance when deployed in natural conversational settings (Imai et al., 2025; Gabler et al., 2023). These challenges are especially pronounced in African linguistic contexts, where multilingualism, dialectal variation, and frequent code-switching are common (Adebara, 2025). Nevertheless, spontaneous speech remains sparsely represented in publicly available African language corpora, limiting systematic investigation of these phenomena. Addressing this gap is essential for developing robust and equitable speech technologies that generalize beyond laboratory conditions.

2.4 Dataset Documentation, Ethics, and Governance

The NLP community has increasingly emphasized the importance of transparent dataset documentation and ethical data practices, as articulated in frameworks such as Datasheets for Datasets (Gebru et al., 2021) and Model Cards (Mitchell et al., 2019). These concerns are particularly salient for speech data, which inherently contains personally identifiable and biometric information. Recent work has further highlighted the societal risks of large-scale language data collection, particularly for underrepresented communities (Bender et al., 2021). In African contexts—where historical data practices have often been extractive—explicit governance frameworks are essential for ensuring responsible data stewardship. The African Voices project builds on these principles by integrating human-in-the-loop anonymization, structured access controls, and lifecycle management into dataset design, positioning governance as a core methodological component rather than a post-hoc consideration. The African Voices 2500 hours corpus addresses these gaps through demographic stratification, diverse speaking contexts, and rigorous quality control. Table 1 compares African Voices with existing speech datasets that include African languages. While prior resources have significantly advanced multilingual speech modeling, they largely emphasize read or scripted speech and provide limited demographic annotation or governance transparency. African Voices addresses these gaps by prioritizing spontaneous speech, rich sociolinguistic metadata, and an explicit access framework.

3 Languages in the Corpus

3.1 Hausa

Hausa is a Chadic language of the Afroasiatic family spoken by more than 80 million people across West and Central Africa. It is characterized by a complex consonant system that includes implosives, ejectives, and labialized or palatalized consonants, as well as a rich vowel inventory with phonemic vowel length. Hausa also employs a tonal system with high, low, and falling tones, although tone is not always represented in standard orthography. These phonological characteristics present unique challenges for speech modeling, making the inclusion of Hausa a critical component of the dataset. Table 3.1 shows the phonemic inventory of Hausa and their orthographic representations.

Lang.	Lang. Family	Subgroup / Branch	# Speakers	Key Linguistic Features
Hausa	Afroasiatic	Chadic (West Chadic)	∼80–90M (L1 + L2)	Tonal; implosive and ejective consonants; vowel length contrast; written in Boko and Ajami scripts
Yoruba	Niger-Congo	Defoid (Yoruboid)	∼40–50M (L1)	Three-level tone system (H, M, L); extensive diacritic use; vowel harmony; rich dialectal variation
Igbo	Niger-Congo	Volta-Niger (Igboid)	∼25–35M (L1)	Tonal; vowel harmony; agglutinative morphology; significant regional lexical variation
Nigerian Pidgin	English-based Creole	Atlantic Pidgin/Creole	∼75–100M (mostly L2)	Non-standardized orthography; simplified morphology; strong influence from Nigerian languages; high code-switching

Table 2: Linguistic and demographic characteristics of the four target languages in the African Voices Project – Nigeria. Population figures are approximate and include both first- and second-language speakers.

Short Vowels	/a/ *a*, /e/ *e*, /i/ *i*, /o/ *o*, /u/ *u*
Long Vowels	/aː/ *aa*, /eː/ *ee*, /iː/ *ii*, /oː/ *oo*, /uː/ *uu*
Consonants	/ɓ/ *b*, /ɗ/ *d*, /ɡ/ *g*, /k/ *k*, /p/ *p*, /t/ *t*, /dʒ/ *j*, /tʃ/ *c*, /ɸ/ *f*, /s/ *s*, /ʃ/ sh, /z/ *z*, /h/ *h*, /m/ *m*, /n/ *n*, /ʔʲ/ 'y, /ʔ/ ', /l/ *l*, /ɾ/ *r*, /ɽ/ *r*, /w/ *w*, /j/ *y*, /kʷ/ *kw*, /kʲ/ *ky*, /gʷ/ *gw*, /gʲ/ *gy*, /kp/ *kp*, /gb/ *gb*, /ɓ/ ɓ, /ɗ/ ɗ, /ƙ/ ƙ. /ʃ/ y

3.2 Igbo

Igbo is another Niger–Congo language with significant internal variation, both phonological and lexical. It features vowel harmony, grammatical tone, reduplication, and rich morphological processes. The project includes data from various regional varieties of Igbo, including Owerri, Nsukka, and Onitsha dialect zones. This diversity ensures that resulting models are not biased toward a single prestige variety. Table 3 summarizes the phoneme inventory alongside their orthographic representations.

Vowels	/a/ *a*, /e/ *e*, /ɛ/ *ẹ*, /i/ *i*, /ɪ/ ɪ, /o/ *o*, /ɔ/ *ọ*, /u/ *u*, /ʊ/ *ụ*
Consonants	/b/ *b*, /kp/ *kp*, /tʃ/ *ch*, /d/ *d*, /f/ *f*, /g/ *g*, /gb/ *gb*, /gh/ *gh*, /gw/ *gw*, /h/ *h*, /dʒ/ *j*, /k/ *k*, /kp/ *kp*, /kw/ *kw*, /l/ *l*, /m/ *m*, /n/ *n*, /ŋ/ *ṅ*, /nw/ *nw*, /ɲ/ *ny*, /p/ *p*, /ɾ/ *r*, /s/ *s*, /ʃ/ *sh*, /t/ *t*, /v/ *v*, /w/ *w*, /j/ *y*, /z/ *z*, /ʒ/ *zh*

Table 3: Igbo phoneme inventory with orthographic representations (in italics)

3.3 Naijá (Nigerian Pidgin)

Naijá (Nigerian Pidgin) is a creole language that functions as a lingua franca across Nigeria. Though rooted in English, it incorporates vocabulary, grammar, and phonological influences from multiple Nigerian languages. It is widely used in informal communication, media, and entertainment. Despite its prevalence, it remains heavily underresourced in formal language technologies. The inclusion of 1000 hours of Nigerian Pidgin speech significantly expands the representational scope of African speech data. In Table 4, we present the phoneme inventory and their corresponding orthographic representation.

Vowels	/a/ *a*, /e/ *e*, /ɛ/ *e*, /ɪ/ *i*, /ʊ/ *u*, /o/ *o*, /ɔ/ *o*
Consonants	/b/ *b*, /p/ *p*, /t/ *t*, /d/ *d*, /f/ *f*, /v/ *v*, /s/ *s*, /z/ *z*, /k/ *k*, /g/ *g*, /gb/ *gb*, /kp/ *kp*, /m/ *m*, /n/ *n*, /l/ *l*, /ɾ/ *r*, /w/ *w*, /j/ *y*, /h/ *h*, /ʃ/ *sh*, /tʃ/ *ch*, /dʒ/ *j*

Table 4: Naijá (Nigerian Pidgin) vowel and consonant inventory, showing IPA representations followed by orthographic symbols (in italics)

3.4 Yoruba

Yoruba belongs to the Niger–Congo language family and is spoken by more than 40 million people. It is a tonal language, using three level tones (high, mid, and low), and meaning can change dramatically based on tonal variation. The dataset was carefully designed to capture this tonal and dialec-

tal diversity, which is essential for building accurate, tone-aware ASR and TTS systems. Table 5 summarizes the phonemic inventory and their orthographic representations.

Oral Vowels	/a/ *a*, /e/ *e*, /ɛ/ *ẹ*, /i/ *i*, /o/ *o*, /ɔ/ *ọ*, /u/ *u*
Nasal Vowels	[ã] *an*, /ɛ̃/ *ẹn*, [ɔ̃] *ọn*, /ĩ/ *in*, /ũ/ *un*
Consonants	/b/ *b*, /d/ *d*, /f/ *f*, /g/ *g*, /gb/ *gb*, /h/ *h*, /dʒ/ *j*, /k/ *k*, /l/ *l*, /m/ *m*, /n/ *n*, /kp/ *p*, /ɾ/ *r*, /s/ *s*, /ʃ/ *ṣ*, /t/ *t*, /w/ *w*, /j/ *y*

Table 5: Yoruba phoneme inventory with orthographic representations (in italics)

Table 2 provides details of the four languages in our corpus.

4 Data Collection

4.1 Dataset Design Goals

The African Voices dataset was designed to support the development and evaluation of automatic speech recognition (ASR) and text-to-speech (TTS) systems for African languages under realistic usage conditions. The primary objective was to collect a large-scale speech corpus that captures *naturally occurring language use*, while remaining sufficiently structured to support reproducible experimentation and fair evaluation. In contrast to datasets that prioritize scripted or read speech, African Voices emphasizes spontaneous and semi-structured speech across domains of high social and economic relevance. The dataset consists of paired audio recordings and orthographic transcripts, accompanied by rich demographic and contextual metadata. All components were systematically organized to support model training, validation, and evaluation across multiple downstream speech tasks.

4.2 Language Scope and Dialectal Considerations

For languages exhibiting substantial internal variation, a reference variety was adopted to ensure orthographic consistency while preserving natural accentual and regional variation in speech. For example, within the Hausa subset, the central dialect (Kananci) was used as the orthographic reference, while speakers representing a wide range of regional accents were included in the recordings. This approach balances linguistic consistency with sociolinguistic realism and avoids collapsing heterogeneous speech communities into a single homogeneous variety.

4.3 Domains and Content Coverage

To ensure real-world applicability, data collection focused on domains central to everyday life and public service delivery, including **agriculture**, **healthcare**, **business**, and **everyday conversation**. Speech data within these domains is particularly important for applications such as agricultural extension services, community health outreach, financial literacy tools, and general-purpose conversational systems. A structured domain taxonomy was developed for each domain to guide data collection and ensure semantic and pragmatic diversity. For instance, the healthcare taxonomy included symptom description, diagnosis, treatment, vaccination, maternal care, and public health messaging, while the agriculture taxonomy covered crop types, planting cycles, pest control, climate variability, irrigation, and market access. Similar taxonomies were developed for financial inclusion and everyday communication. These taxonomies served as blueprints to avoid narrow or repetitive content and to ensure broad thematic coverage across recording sessions. Figure 1 illustrates the taxonomy of the health domain and the relative prominence of its major thematic categories while the breakdown of domains in the dataset is presented in Figure 2f.

4.4 Speech Elicitation and Recording Protocol

Speech was elicited using a combination of multimodal prompts, including text-based, image-based, and audiovisual stimuli. Each modality was selected to encourage different types of linguistic production. Text prompts were primarily used for controlled or read speech, while visual and audiovisual prompts were particularly effective in eliciting spontaneous and semi-spontaneous responses. For spontaneous speech collection, participants were presented with prompts depicting realistic scenarios—such as images of farming activities, short videos simulating medical consultations, or visual representations of mobile money transactions—and were asked to respond or narrate freely in their own words without written scripts. This approach captures natural hesitations, reformulations, discourse markers, and pragmatic phenomena that are typically absent from read speech. For read speech, participants were provided with carefully designed textual materials that were phonetically, syntactically, and lexically rich. Source texts were

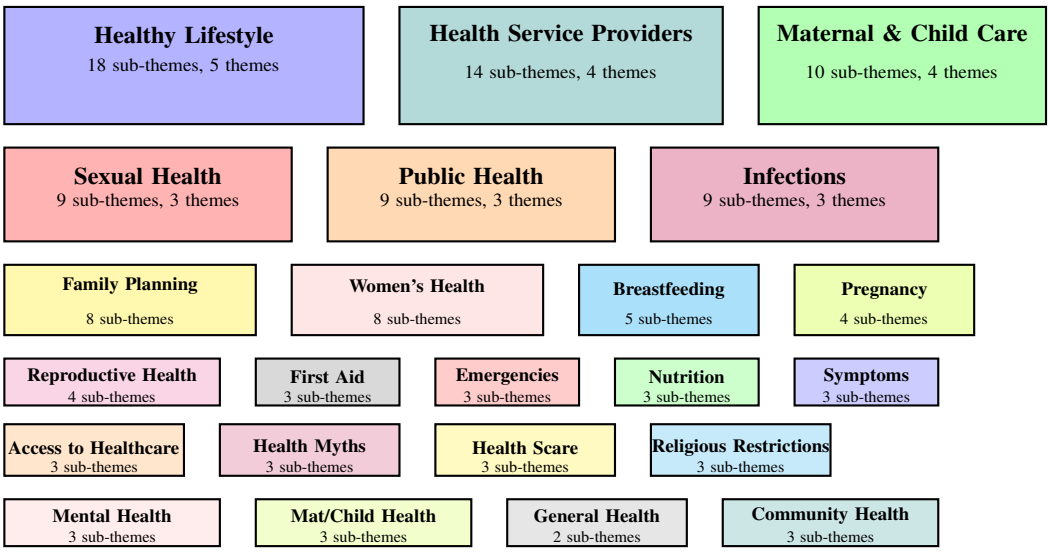


Figure 1: Treemap visualization of health domain taxonomy showing relative importance of each dominant engagement point. Larger boxes indicate more coverage with higher sub-theme counts. Healthy Lifestyle (18), Health Service Providers (14), and Maternal & Child Care (10) dominate the framework.

either authored directly in the target language or translated from English by professionally trained translators working in collaboration with native speaker consultants. Particular care was taken to ensure cultural appropriateness and to avoid literal translations that could introduce unnatural phrasing. Recordings were conducted in quiet, naturalistic environments using commonly available devices, including smartphones and portable microphones. This variability in recording conditions was intentional and reflects realistic deployment scenarios for speech technologies.

4.5 Speaker Recruitment and Metadata

All recordings were produced by native speakers of the target languages. Speakers were recruited through community-based networks and local institutions to ensure diversity across age, gender, educational background, and geographic region. To maximize speaker diversity, individual contributors were typically limited to approximately one hour of recorded speech per language. In total, the dataset includes contributions from 2,865 unique speakers.

For each recording, structured metadata was collected, including:

- **Age group**: e.g., 15–29, 30–45, 46–60, 61+;
- **Gender**: self-identified;
- **Educational qualification**: e.g., no formal education, primary, secondary, tertiary.

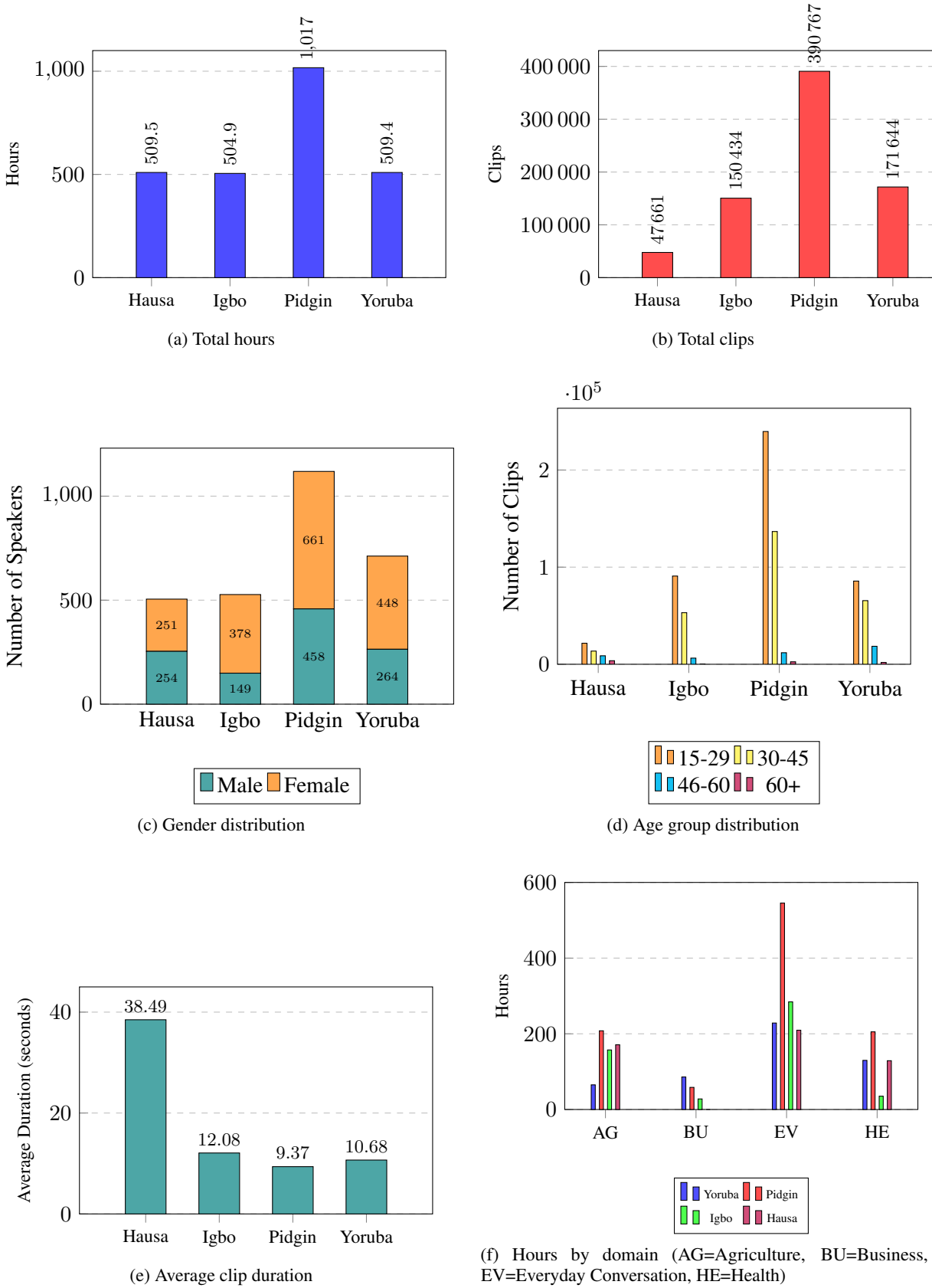
This metadata supports sociolinguistic analysis, fairness-aware evaluation, and controlled experimental design. Figure 2d shows the distribution of speaker's ages across each language in the dataset. Gender coverage in represented in Figure 2c.

4.6 Audio Quality Control

All recordings underwent a dual-layer quality assurance process. Automatic checks verified technical criteria, including signal-to-noise ratio (targeting approximately 40 dB) (ITU-T, 2018), sampling rate (48 kHz) (Audio Engineering Society, 2019), and bit depth. In parallel, manual review was conducted by trained native speakers who verified that recordings contained authentic speech in the correct language, were free from excessive background noise, and were representative of native speaker usage. Only recordings that passed both automatic and manual checks were included in the final dataset.

4.7 Transcription Process

All spontaneous speech recordings were transcribed by native speakers using a standardized set of transcription guidelines developed by professional linguists. These guidelines specify orthographic conventions, use of diacritics, treatment of disfluencies, and annotation of code-switching. Disfluencies such as hesitations and repetitions are



(a) Total hours

(b) Total clips

(c) Gender distribution

(d) Age group distribution

(e) Average clip duration

(f) Hours by domain (AG=Agriculture, BU=Business, EV=Everyday Conversation, HE=Health)

Figure 2: Comprehensive analysis of Nigerian language speech dataset. (a-b) Dataset size metrics: total recording hours and number of audio clips showing Pidgin's dominance; (c) Gender distribution with male speakers outnumbering female across all languages; (d) Age group distribution revealing varying demographic patterns; (e) Average clip duration demonstrating consistency across languages; (f) Domain distribution showing language specialization across different subject areas.

marked using the placeholder tag **[um]**, unintelligible segments are marked as **[?]**, and instances of code-mixing are enclosed using the **[cs]** tag. Numbers and special characters are spelled out in full to support downstream speech processing tasks. Transcribers were instructed to transcribe speech verbatim, preserving grammatical errors and informal constructions as spoken. Each transcript was reviewed by linguists with both native-speaker fluency and formal linguistic training, enabling consistent handling of dialectal variation and subtle transcription errors.

4.8 Dataset Splits

To support fair evaluation and prevent speaker leakage, the dataset was partitioned using a speaker-level stratified split. Stratification was performed based on language, age group, gender, and recording type. Speakers were uniquely identified and assigned exclusively to one of four splits: **train** (approximately 85%), **development** (approximately 5%), **development-test** (approximately 5%), and **test** (approximately 5%). For strata with insufficient representation, speakers were assigned randomly according to predefined proportions. After splitting, integrity checks verified that no speaker appeared in more than one split, all clips were assigned, and demographic balance was preserved. This procedure yields a reproducible and balanced partitioning suitable for robust model evaluation.

5 Experimental Setup

We conducted preliminary experiments using two Automatic Speech Recognition (ASR) models based on the Whisper architecture: Whisper-Small and Whisper-Large. All models were fine-tuned under identical training conditions to ensure a fair comparison across model sizes and languages. Training was performed for a maximum of 10 epochs with a batch size of 32 per GPU. We used a learning rate of 1e-5 with a warmup ratio of 0.1, and enabled mixed-precision (FP16) training to improve computational efficiency and reduce memory usage. Optimization settings were kept constant across all experiments. To accelerate training, we employed distributed data-parallel training using PyTorch's Distributed Data Parallel (DDP) framework, running experiments on 3–4 GPUs depending on availability. To mitigate overfitting, early stopping was applied with a patience of 3 epochs, based on performance on the development set. Model checkpoints were evaluated using Word Error Rate (WER), which served as the primary evaluation metric throughout training and model selection. Table 6 shows performance across all languages.

Language	Small	Turbo	Large	Omni*
Hausa	19.86	12.81	**12.57**	36.00
Igbo	52.95	45.58	**44.49**	58.00
Naija	9.98	10.29	**8.28**	44.00
Yoruba	38.92	33.59	**30.02**	44.00

Table 6: WER comparison of fine-tuned Whisper models and Omni* (zero-shot).

6 Conclusion

We have presented African Voices, a 2,500-hour speech dataset spanning four Nigerian languages (Hausa, Igbo, Naija Pidgin, and Yoruba) with 2,865 speakers, emphasizing spontaneous and semi-structured speech that captures real-world linguistic variation. Our baseline ASR experiments demonstrate substantial improvements from fine-tuning. The dataset's rich metadata and ethical governance framework position it as a foundational resource for advancing speech technology research on African languages. Future work will focus on three key areas: (1) fine-grained error analysis to understand how models handle tonal contrasts, morphological complexity, and code-switching; (2) systematic investigation of model behavior across demographic groups and recording conditions to identify and mitigate performance biases; and (3) exploration of architectural innovations and training strategies tailored to the linguistic characteristics of African languages, including multilingual pre-training, tone-aware modeling, and linguistically informed tokenization. We will also conduct comparative analyses across languages to distinguish universal low-resource ASR challenges from language-specific phenomena. We release African Voices to enable rigorous empirical investigation and contribute to more inclusive speech AI.

References

Firdus Abu, Dawit Kebede, and Getachew Tulu. 2025. Sagalee: The first 100-hour oromo automatic speech recognition corpus. In *Proceedings of ACL 2025*.

Martine Adda-Decker, Gilles Adda, and Lori Lamel. 2000. Investigating text normalization and pronunciation variants for german broadcast transcription. In

6th International Conference on Spoken Language Processing (ICSLP 2000), pages vol. 1, 266–269.

Ife Adebara. 2025. Ai and language data flaring in africa: Addressing the low-resource challenge. Technical report, Centre for International Governance Innovation.

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. 2020a. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020b. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Audio Engineering Society. 2019. Aes standard for digital audio engineering.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens. 2007. Automatic speech recognition and speech variability: A review. *Speech Communication*, 49(10):763–786. Intrinsic Speech Variations.

Laurent Besacier, Elodie Gauthier, Mathieu Mangeot, Philippe Bretier, Paul Bagshaw, Olivier Rosec, Thierry Moudenc, François Pellegrino, Sylvie Voisin, Egidio Marsico, and Pascal Nocera. 2015. Speech Technologies for African Languages: Example of a Multilingual Calculator for Education. In *Interspeech 2015 (short demo paper)*, Dresden, Germany.

Alexis Conneau, Myle Ott Ma, Kushal Lakhotia, and Luke Zettlemoyer. 2022. FLEURS: Few-shot learning evaluation of universal representations of speech. In *ICASSP 2022*, pages 6523–6527.

Chris Emezue, NaijaVoices Community, Busayo Awobade, Abraham Owodunni, Handel Emezue, Gloria Monica Tobechukwu Emezue, Nefertiti Nneoma Emezue, Sewade Ogun, Bunmi Akinremi, David Ifeoluwa Adelani, and 1 others. 2025. The naijavoices dataset: Cultivating large-scale, high-quality, culturally-rich speech data for african languages. *arXiv preprint arXiv:2505.20564*.

Philipp Gabler, Bernhard C. Geiger, Barbara Schuppler, and Roman Kern. 2023. Reconsidering read and spontaneous speech: Causal perspectives on the generation of training data for automatic speech recognition. *Information*, 14(2).

Elodie Gauthier, Laurent Besacier, Sylvie Voisin, Michael Melese, and Uriel Pascal Elingui. 2016. Collecting resources in sub-saharan african languages for automatic speech recognition: a case study of wolof. *LREC*.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM*, 64(12):86–92.

Hadrien Gelas, Laurent Besacier, and Francois Pellegrino. 2012. Developments of Swahili resources for an automatic speech recognition system. In *SLTU - Workshop on Spoken Language Technologies for Under-Resourced Languages*, Cape-Town, Afrique Du Sud.

J.J. Godfrey, E.C. Holliman, and J. McDaniel. 1992. Switchboard: telephone speech corpus for research and development. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520 vol.1.

Saki Imai, Tahiya Chowdhury, and Amanda J. Stent. 2025. Evaluating open-source ASR systems: Performance across diverse audio conditions and error correction methods. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5027–5039, Abu Dhabi, UAE. Association for Computational Linguistics.

ITU-T. 2018. Objective measurement of active speech level. (P.56).

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

Alexis Lau, Adedayo Fayemi, and Nkemdilim Uche. 2025. Data quality issues in multilingual speech datasets. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Florian Meyer, Henk Heuvel, and Amarsanaa Uyanga. 2022. BibleTTS: Studio-quality single-speaker speech for ten african languages. In *Proceedings of Interspeech 2022*.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA. Association for Computing Machinery.

Masanobu Nakamura, Koji Iwano, and Sadaoki Furui. 2008. Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. *Computer Speech Language*, 22(2):171–184.

Emmanuel Ogun, Jiyoon Choi, Bonaventure F. Dossou, and Jesujoba Alabi. 2024. Afrotts: Generating 86 african english accents with 747 voices. In *Interspeech 2024*.

Rotimi Ogunremi, Fisayo Soetan, and Oladipupo Awobode. 2023. ÌròyìnSpeech: Building a multi-domain yoruba speech corpus. In *Proceedings of Interspeech 2023*.

Olaniyi Olatunji, Jesujoba Alabi, Bonaventure F. Dossou, David Adelani, and Jade Abbott. 2023. AfriSpeech-200: A pan-african accented english speech dataset. In *Transactions of the Association for Computational Linguistics*, volume 11, pages 3187–3204.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Martha Tachbelie, Solomon Teferra Abate, and Laurent Besacier. 2014. Using different acoustic, lexical and language modeling units for asr of an under-resourced language - amharic. *Speech Communication*, 56.

Omnilingual ASR team, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, Chierh Cheng, Joe Chuang, Caley Droof, Mark Duppenthaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, and 14 others. 2025. Omnilingual asr: Open-source multilingual speech recognition for 1600+ languages. *Preprint*, arXiv:2511.09690.