# AfriBase-Speech: Auditing, Tiering, and Packaging African Speech Resources from Government Public Domain Archives

**Ernst van Gassen**
Independent researcher at Arktos AI Lab
The Netherlands
evg.gassen@gmail.com
ORCID 0009-0000-0314-6441

## Abstract

African speech data for natural language processing (NLP) has grown rapidly, but the licensing that governs it is fragmenting. We document *license washing* in four African speech datasets: a permissive Creative Commons Attribution (CC BY) tag on the hosting platform, contradicted by restrictive terms in the data card, end-user license agreement, or terms of service. As an alternative we package source audio that needs no license in the United States (US) into a dataset released under CC BY 4.0. The US Foreign Service Institute, Peace Corps, and Defense Language Institute produced language courses to train diplomats, volunteers, and service members, and recorded the lessons of those courses as audio ("lesson audio"). As works of the US federal government, these recordings are in the public domain in the US under Section 105 of the US copyright law (Title 17 of the US Code), yet they have never been prepared for NLP use. We inventory such lesson audio for 22 African languages and Haitian Creole. Our pipeline splits each recording into utterances, labels them with Whisper large-v3, and removes the English speech of the course instructor by language identification; English makes up 2 to 36% of segments, depending on the language. Fine-tuning Whisper-small on the result lowers word error rate on the FLEURS (Few-shot Learning Evaluation of Universal Representations of Speech) benchmark for three of five languages, by 13 to 57 points. Removing English moves it by less than 2 points except for Amharic, where the quality of the machine labels limits training. The corpus, its per-language datasheets, and the pipeline code are released.

**Keywords:** speech data licensing, license washing, public domain audio, African languages, corpus provenance, Whisper fine-tuning

## 1 Introduction

Over the past five years African speech resources for natural language processing (NLP) have grown substantially. Corpora covering dozens of languages are indexed on HuggingFace, Zenodo, and OpenSLR (three public data-hosting platforms), many with Creative Commons (CC) tags that signal permissive use. Open licensing was designed for exactly this legibility: a small vocabulary of standard tags that communicate use rights at a glance. The tag used most often is CC BY, the Creative Commons Attribution license, which permits any use, including commercial use and modification, provided the source is credited.

In a growing number of cases, the legibility misleads. A CC BY 4.0 tag (version 4.0 of that license) records what the depositor *claims* about use rights, not what a downstream user actually acquires. The depositor's own end-user license agreement (EULA), data card, or terms of service may impose restrictions that the tag gives no sign of and that contradict the freedoms a CC BY license grants. Examples include bans on speech synthesis, restrictions on commercial artificial intelligence (AI) use, and content the depositor does not own. Anyone who relies on the tag without reading the underlying documents inherits the error.

We call this pattern *license washing*: a permissive tag that implies broader rights than the actual terms provide. It need not be deliberate; it often reflects uncertainty about what a CC license can cover in an aggregated collection. Its effect does not depend on intent. A practitioner who checks compliance by scraping metadata will miss undocumented restrictions at scale.

This paper studies two kinds of material. The first is a screen of 20 African speech datasets. Four show license washing: DDD-Kenya, a 1,664-hour corpus of Luhya, Kamba, Gusii, and Somali speech; Swivuriso, a South African corpus; the

This work is licensed under CC BY SA 4.0. To view a copy of this license, visit https://creativecommons.org/licenses/by-sa/4.0/deed.en
The copyright remains with the authors.

West African Radio Corpus (SLR105), 142 hours of West African radio audio; and Gamayun Kenyan Swahili, a corpus distributed by CLEAR Global. Two more, Esethu (isiXhosa speech from Lelapa AI) and the DhoNam Dholuo corpus, which uses the Nwulite Obodo Open Data License (NOODL), carry intentional community-protective licenses, which are legitimate data sovereignty choices rather than washing. The second kind of material is language-course audio from three US federal agencies, the Foreign Service Institute (FSI), the Peace Corps, and the Defense Language Institute (DLI): lesson recordings for 22 African languages and Haitian Creole, in the public domain (PD) in the US. We segment this audio, label it automatically with the Whisper large-v3 speech recognition model, remove the English speech of the course instructors, and package the result. We group material into tiers by the legal basis of its rights status. Tier 1 is material in the public domain because US law denies copyright to works of the federal government. The lesson audio is Tier 1a. Voice of America (VOA) broadcast audio, also US government work, is Tier 1b, described in a companion paper (van Gassen, 2026b) accepted at this conference and to appear in the same proceedings. We evaluate the lesson audio by fine-tuning the Whisper-small model for five languages and measuring word error rate on FLEURS (Few-shot Learning Evaluation of Universal Representations of Speech), a read-speech benchmark covering more than 100 languages.

This paper makes three contributions:

1. **License audit.** Of 20 African speech datasets screened, we document license washing in four, and distinguish them from two intentional community-protective licenses, Esethu and NOODL.
2. **Public domain corpus.** We inventory, download, segment, and package the US government PD lesson audio for 22 African languages and Haitian Creole. To our knowledge this is the first systematic preparation of this material for NLP.
3. **Downstream evaluation.** We fine-tune Whisper-small on the segmented audio for five languages and measure word error rate on FLEURS.

The rest of the paper is organised as follows. Section 2 reviews existing African speech resources, licensing frameworks, and government public domain material. Section 3 presents the license audit. Section 4 describes the lesson audio and the processing pipeline, and Section 5 the Voice of America tier. Section 6 reports the evaluation, Section 7 the released package, and Sections 8 and 9 give the discussion and conclusion. Limitations follow the conclusion. Appendix A gives the lesson audio inventory and Appendix B the screened datasets.

## 2 Background

### 2.1 African speech resources and their gaps

The FLEURS benchmark (Conneau et al., 2023) covers 102 languages under CC BY 4.0, more than 20 of them African, and has become the standard evaluation set for low-resource speech recognition. Meta's Massively Multilingual Speech (MMS) project extended automatic speech recognition (ASR) to 1,107 languages (Pratap et al., 2024). These models have outpaced the supply of training data with clear licensing for the languages they cover.

The corpus of the National Centre for Human Language Technology (NCHLT) (Barnard et al., 2014) provides 56 hours per language for the 11 official languages of South Africa under CC BY, and remains the main labeled resource for them. A lesser known follow-up, the NCHLT Auxiliary corpora, adds 1,421 hours under CC BY 4.0 (Badenhorst and de Wet, 2022); released in 2019, it postdates the training of major foundation models and has seen limited uptake. WAXAL (Diack et al., 2026) is the largest recent addition, with about 1,250 hours of image-prompted speech in 24 African languages and a separate 235-hour single-speaker text-to-speech (TTS) component. Its paper states a uniform CC BY 4.0 license, but its HuggingFace card does not, as van Gassen (2026a) showed. Only the University of Ghana contributions (Akan, Ewe, Dagbani, Dagaare, Ikposo) are CC BY 4.0; the rest, including Lingala, Hausa, Igbo, and Yoruba, are CC BY-SA 4.0. CC BY-SA is CC BY plus a share-alike (SA) condition: anything built from the material must be shared under the same license. A blanket CC BY assumption therefore does not hold for most of the corpus.

No existing resource treats US government language training audio systematically. FSI, the Peace Corps, and DLI produced African language courses from the 1960s onwards. The courses circulate as individual audio files in the MP3 format and as textbooks in the Portable Document

Format (PDF). The Live Lingua Project indexes them by agency, with the African courses at paths such as `livelingua.com/project/fsi/yoruba` and `livelingua.com/project/peace-corps/ewe`, and the Internet Archive holds mirrored copies (both accessed 17 October 2026). We downloaded the audio from the Internet Archive; the item identifiers ship with the corpus. The Archive's terms of use grant access "for scholarship and research purposes only" (Internet Archive, 2014). Our use is research. These terms govern access to the Archive and do not change the public domain status of the recordings. Neither host offers annotation or NLP packaging, and no prior work has segmented or evaluated the material.

### 2.2 Licensing frameworks and their limits

Creative Commons licenses were designed for creative and scholarly works, and three of their provisions cause friction in computational use. The noncommercial (NC) restriction is deliberately broad, and what it covers in computational use is unsettled. The SA obligation raises open questions for derived datasets and model weights. Most important here is scope: a CC BY license on a database does not cover content whose copyright holders are not party to it.

Some recent African speech datasets instead adopt community-protective licenses: Esethu for Lelapa AI's isiXhosa dataset (Rajab et al., 2025), and NOODL for the DhoNam Dholuo corpus (Okorie and Omino, 2025). These respond legitimately to extractive data practices and are distinct from license washing (Section 3.3).

### 2.3 Government public domain in NLP

LibriSpeech built a large speech corpus from public domain audiobooks (Panayotov et al., 2015), and Common Voice added crowdsourced recordings under CC0 (Ardila et al., 2020), a public-domain dedication by which the contributors waive their rights as far as the law allows. Neither addresses government language-instruction audio. Under Section 105 of Title 17 of the US Code (17 U.S.C. § 105), US copyright protection is not available for works prepared by US government officers or employees as part of their official duties. This covers the courses that FSI, Peace Corps, and DLI staff produced: recordings, textbooks, vocabulary lists, and transcripts, regardless of language or date.

## 3 License Washing: Documented Cases

### 3.1 Audit method

We screened 20 African speech datasets that a practitioner building a speech recognition or speech synthesis pipeline is likely to encounter: those cited in recent African speech papers or listed on OpenSLR, HuggingFace, and Zenodo under an African language. Appendix B lists all 20 with our finding for each. For each dataset we first retrieved the license tag from the hosting platform (the HuggingFace card, Zenodo record, or OpenSLR page). We then read the full EULA, data card, README, and platform terms, and recorded any restriction that contradicts or narrows the tag. We classify a discrepancy as *washing* when the tag misrepresents the actual terms, and as an *intentional restriction* when the terms are stated clearly but are not a standard CC license. The first is a documentation failure; the second is a deliberate licensing choice. Of the 20, four show washing, two carry intentional restrictions, and the rest are either correctly tagged or out of scope.

### 3.2 Washing cases

**DDD-Kenya.** This 1,664-hour corpus of Luhya, Kamba, Gusii, and Somali speech carries a CC BY 4.0 tag in its HuggingFace metadata. Its EULA prohibits "use in commercial artificial intelligence systems," "voice cloning," and "biometric identification." CC BY 4.0 does not allow such restrictions to be added to licensed material, yet the EULA may still bind users as a contract regardless of the tag.

**Swivuriso (za-african-next-voices).** This South African corpus (arXiv 2512.02201) carries a CC BY 4.0 tag on HuggingFace. Access is gated by a click-through EULA that bans "any form of text-to-speech (TTS), voice cloning, voice synthesis" and records that the speakers "explicitly withhold consent for voice synthesis applications." A user who reads only the repository metadata never sees the restriction. We read the EULA by passing the access gate. The card publishes no corpus size, so we report none.

**West African Radio Corpus (SLR105).** This 142-hour corpus of West African radio audio is released under CC BY-SA 4.0. Radio broadcasts contain music, news wire segments, and other third-party content whose copyright the depositors do not hold, so they cannot license it to others under any CC license. The CC license therefore does not cover that content.

| Dataset | Tag | Actual terms | Consequence |
|---|---|---|---|
| DDD-Kenya (1,664h) | CC BY 4.0 | EULA bans commercial AI, voice cloning | commercial AI use not permitted |
| Swivuriso | CC BY 4.0 | gated EULA bans TTS and voice cloning; speakers withhold consent | TTS use not permitted |
| SLR105 radio (142h) | CC BY-SA 4.0 | third-party music and news the depositors cannot license | rights in third-party content not cleared |
| Gamayun Swahili (6h) | "open" | non-commercial only | commercial use not permitted |

Table 1: License washing cases in African speech datasets.

**Gamayun Kenyan Swahili.** CLEAR Global distributes this corpus under a custom license. It is described as "open," but the license restricts use to non-commercial purposes, and repository descriptions do not surface the restriction.

Table 1 summarises the four cases.

### 3.3 Intentional community-protective licenses

Two datasets impose restrictions that are deliberate and clearly stated. We include them because practitioners may meet them as apparent CC alternatives, but we do not classify them as washing: the restrictions are the point.

**Esethu (isiXhosa, 10 hours, Lelapa AI).** The Esethu license requires commercial users in countries of the Organisation for Economic Co-operation and Development (OECD) to share benefits, so that the communities whose speech is recorded receive value from commercial use. The license is clearly documented; the problem is that a platform may surface only a CC-style label.

**NOODL (DhoNam Dholuo, 51 hours).** NOODL tiers access by geography: users in the Global South may use the data without royalty, while commercial users in OECD countries must agree to benefit sharing. It moves licensing from restricting copies towards restricting uses and sharing benefits, which deserves attention whether or not it fits the CC categories.

Both continue the participatory tradition of the Masakhane community (Nekoto et al., 2020). AfriBase-Speech uses neither dataset, not because the frameworks are wrong, but because our design goal is a corpus that anyone may redistribute with attribution as the only condition.

### 3.4 Implication

Scraping metadata is not enough for license compliance in African language NLP. Any dataset used in a redistributable pipeline needs a manual reading of its actual terms. AfriBase-Speech is designed to make that reading unnecessary: its sources are in the public domain by US statute and carry no hidden terms, and the compiled dataset carries a single standard license, CC BY 4.0.

## 4 Tier 1a: Government Lesson Audio

### 4.1 Legal basis

FSI (part of the State Department), the Peace Corps, and DLI (part of the Department of Defense) are US federal agencies, and their staff recorded the lesson audio as part of official language training. Under Section 105 these works carry no copyright in the US, so they need no license there and no negotiation. We nonetheless attribute every recording to its agency and course. US copyright law does not require this, but public domain status does not end every claim to be named: in some countries, such as France, the author's moral right to be identified does not expire. Attribution also keeps the provenance of each recording checkable.

### 4.2 Source inventory

We identified accessible PD lesson audio for 22 African languages and Haitian Creole across the three agencies. Volume ranges from Amharic (180 files, 21.2 hours, FSI) to SiSwati (9 files, about 45 minutes, Peace Corps). Five languages, Amharic, Hausa, Lingala, Swahili, and Yoruba, combine enough audio with a FLEURS test set; they are the evaluation languages of Section 6. Appendix A gives the full inventory.

Nine SiSwati files are a thin resource, but SiSwati is a national language of Eswatini, which hosts this conference. We include it regardless of volume.

The material exists only as course-level MP3 files, without segmentation, transcripts aligned to the audio, or utterance-level metadata.

### 4.3 The mixed-language problem

In lesson audio, an English-speaking instructor introduces phrases, explains grammar, and translates, while a native speaker produces the target-language

material slowly and clearly. The two voices alternate throughout. Audio ingested without filtering therefore mixes English instruction into data intended for another language. The English share varies with course design and position. In the Yoruba course it falls from 59.3% of segments in the first half of the files to 28.1% in the second, while Swahili shows no such decline. Section 4.5 measures it per language.

### 4.4 Pipeline

The pipeline has four stages, applied per language. Its components are established practice: long recordings are routinely split at pauses into utterances, and labelling untranscribed speech with a larger recogniser is the approach used at scale to train Distil-Whisper (Gandhi et al., 2023). What is specific to this pipeline is the removal of the instructor's speech in a second language.

**Segmentation.** Each file is converted to 16 kHz mono and split at pauses, using silence detection in the pydub library[1] (pauses of at least 500 milliseconds below −35 decibels relative to full scale). Lesson audio leaves a pause after each phrase for the student to repeat, so these splits recover utterance boundaries.

**Quality filtering.** A segment is kept if it lasts 1 to 15 seconds, has a signal-to-noise ratio (SNR) of at least 10 decibels (dB), and has a speech fraction of at least 0.60. SNR is estimated from frame energy, comparing the loudest quarter of frames with the quietest quarter; speech fraction is the share of the segment above the silence threshold.

**Pseudo-labeling.** The source courses have printed textbooks but no transcripts aligned to the audio, so each segment is transcribed by Whisper large-v3 (Radford et al., 2023) with the target language set in advance. Setting the language matters: with automatic detection, 98.5% of Amharic labels come out in Arabic rather than Ge'ez script, against 5.0% when the language is set. The labels are therefore machine output, not human transcriptions, and their quality is bounded by large-v3's ability in each language (Section 6).

**Language identification.** Each segment is passed to Whisper large-v3's language identifier through the faster-whisper implementation.[2] A segment whose most likely language is English is removed as instructor speech. This step is fully automatic: the identifier's decisions were not checked by listening or against human language labels (see Limitations). We do not require the detected language to match the target, because detection is unreliable on short segments of low-resource languages; the instruction language, by contrast, is consistently recognised.

The pipeline code is released (Section 7).

[1] https://github.com/jiaaro/pydub

[2] https://github.com/SYSTRAN/faster-whisper

| Language | Segments | English | Kept | Kept hours |
|---|---|---|---|---|
| Amharic | 9,442 | 12.9% | 8,225 | 5.23 |
| Hausa | 1,080 | 15.4% | 914 | 0.38 |
| Lingala | 3,430 | 2.1% | 3,359 | 1.69 |
| Swahili | 4,245 | 13.9% | 3,656 | 1.57 |
| Yoruba | 3,943 | 36.3% | 2,510 | 1.40 |

Table 2: Effect of language identification on the evaluation languages (alphabetical). "Segments" passed the quality filter; "English" is the share removed as instructor speech; "Kept" and "Kept hours" are what remains.

### 4.5 How much of the audio is English

Table 2 reports the language identification stage. English instructor speech accounts for 2.1 to 36.3% of segments and 1.5 to 41.5% of segment duration, depending on the language. The share reflects how much of each course is taught in English, from 41.5% of Yoruba duration to 1.5% of Lingala duration. The detector rarely names the target language for the kept segments: it labels 56.8% of kept Amharic segments as Arabic and only 19.9% of kept Yoruba segments as Yoruba, which is why we filter on English alone. The counts in this table, not the raw course duration, are the training data used in Section 6.

### 4.6 Properties of the audio

Lesson audio is useful but limited as ASR training data. Speakers were chosen for clear pronunciation and speak slowly. On the other hand, the domain is narrow (greetings, travel, government services, basic commerce), each course has one to three speakers, and the register is pedagogical rather than conversational. Models trained on it may not transfer to conversational or broadcast speech, a question the companion paper takes up (van Gassen, 2026b).

## 5 Tier 1b: Voice of America Broadcast Audio

Voice of America (VOA) is operated by the US Agency for Global Media, a federal agency, and

VOA's terms of use state that material it produces is in the public domain. After VOA's Africa division closed in March 2025, its archive remains online but is no longer maintained. We include VOA audio as a second PD tier because it offers what lesson audio lacks: natural, news-paced speech from many native speakers.

The two tiers differ acoustically. Lesson audio is clean, slow, and studio-recorded. VOA audio is radio-compressed, involves many speakers, and is delivered at natural pace under varying noise conditions. Without reference transcripts it is better suited to self-supervised pretraining and domain adaptation than to supervised training. Kimanuka et al. (2023) released 741 hours of unlabeled Congolese radio under CC BY, and pretraining on it lowered the Lingala word error rate (WER) from 13.3% for a time-delay neural network baseline to 6.81%. That corpus can no longer be rebuilt from its sources, however: 402 of its 405 source videos are no longer online (van Gassen, 2026a). VOA Lingala offers a public domain counterpart.

VOA broadcasts also contain material VOA does not own: music between segments, wire service clips, and outside interviewees. Section 105 does not cover these. The companion paper (van Gassen, 2026b) describes how the archive is crawled, segmented, filtered by SNR and page language, and transcribed; we package its output as Tier 1b. Lingala, Hausa, Amharic, Somali, Shona, and Afaan Oromo have both VOA audio and a FLEURS test set, and are the natural targets for pretraining on this tier, which we leave to future work.

## 6 Downstream Evaluation

**Question.** Does fine-tuning Whisper-small on Tier 1a lesson audio lower its word error rate in languages where its zero-shot performance is weak?

**Metric.** WER counts the substitutions, deletions, and insertions needed to turn the system output into the reference, divided by the number of reference words. Because insertions count, WER exceeds 100% when a system outputs more wrong words than the reference contains, which is common for low-resource languages.

**Setup.** We compare Whisper-small (244 million parameters, released under the MIT license, a permissive software license that allows any use, modification, and redistribution provided the copyright notice is kept) zero-shot with the same model fine-tuned on the language's Tier 1a segments, both evaluated on the FLEURS test split with the target language set in advance. Each language is trained separately. The segments of Table 2 are split 95/5 into training and validation sets. Training runs for 10 epochs with batch size 8, learning rate $10^{-5}$, warm-up of at most 200 steps, and mixed precision; we keep the epoch checkpoint with the lowest validation WER against the pseudo-labels. Training used a single laptop graphics processing unit (GPU; NVIDIA RTX 5090 Laptop, 24 GB) and took between 16 minutes (Hausa) and 3.1 hours (Amharic) per language. No computing cluster or cloud budget is needed: a researcher or a small group with one consumer machine can repeat the fine-tuning, or apply it to another language. To isolate the language identification stage, we also report the same training on the segments before English was removed.

**Results.** Table 3 gives the results; lower WER is better. Fine-tuning on the filtered segments helps for three languages: WER falls by 13.4 points for Lingala, 22.3 for Swahili, and 56.7 for Yoruba, although Yoruba remains above 100%. Hausa is essentially unchanged, and Amharic gets worse by 60.7 points.

Does removing the English instructor speech matter for these scores? For four languages it changes WER by less than 2 points (between $-1.7$ and $+1.4$) compared with training on the unfiltered segments; for Amharic it changes WER by $+9.5$ points. For Yoruba this holds even though the filter removes 41.5% of the training audio: 1.3 hours of Yoruba give the same result as 2.3 hours of mixed Yoruba and English. On this benchmark, then, the filter changes what the corpus contains more than how well the model scores. Each condition was run once, so differences of one or two points should not be read as effects.

**Why some languages do not improve.** Training volume does not predict the outcome: Amharic has by far the most segments yet regresses most. The quality of the pseudo-labels is the more plausible driver. They come from Whisper large-v3, whose own zero-shot FLEURS WER is 157.9% for Amharic and 86.2% for Hausa, against 35.2% for Swahili (van Gassen, 2026b). Where the labeling model cannot transcribe a language, fine-tuning teaches the small model its errors. A second factor is domain mismatch between slow pedagogical speech and FLEURS's read news sentences, which affects all five languages alike and so cannot explain why the direction differs.

| Language | FLEURS | Train hours | Zero-shot WER | Fine-tuned WER | | Change vs. zero-shot |
|---|---|---|---|---|---|---|
| | | | | without LangID | with LangID | |
| Amharic | am_et | 4.97 | 115.7 | 166.9 | 176.4 | +60.7 |
| Hausa | ha_ng | 0.36 | 90.6 | 91.2 | 91.2 | +0.6 |
| Lingala | ln_cd | 1.61 | 89.9 | 78.2 | 76.5 | −13.4 |
| Swahili | sw_ke | 1.50 | 73.2 | 52.1 | 50.9 | −22.3 |
| Yoruba | yo_ng | 1.33 | 166.0 | 107.9 | 109.3 | −56.7 |

Table 3: Whisper-small WER (%) on the FLEURS test split, languages in alphabetical order. LangID stands for language identification. "Without LangID" trains on all quality-filtered segments, including English instructor speech; "with LangID" trains on the segments kept in Table 2, whose training split is given as "Train hours". Change compares the with-LangID model to zero-shot, in percentage points; negative is better. Lingala, Swahili, and Yoruba have mean segment SNR of 14.9, 17.3, and 19.3 dB; Amharic and Hausa have 18.3 and 16.3 dB.

## 7 The AfriBase-Speech Corpus Package

### 7.1 Format

Each language is distributed as a directory of 16 kHz mono audio files in the Waveform Audio (WAV) format with a metadata file in the JSON Lines (JSONL) format, one record per utterance. Language codes follow ISO 639-3, the three-letter standard of the International Organization for Standardization (here `yor` for Yoruba; the other evaluation languages are `swh` for Swahili and `lin` for Lingala). A record looks like this:

```
{"utterance_id": "yor_u001442",
 "language": "yor", "source_type": "FSI",
 "source_file": "yrb03-19-01.mp3",
 "license": "PD-US",
 "license_basis": "17 U.S.C. § 105",
 "audio_path": "yor/yor_u001442.wav",
 "duration_s": 4.19, "snr_db": 26.7,
 "speech_fraction": 0.666,
 "span_start_s": 191.4, "span_end_s": 195.59,
 "text": "E lo dukwe lo wo babayin.",
 "text_source": "whisper-large-v3-pseudo",
 "langid_top": "yo", "langid_prob": 0.71}
```

The `text_source` field marks every label as machine-generated, and the language identification fields let users apply a stricter filter of their own. Like 98.0% of the kept Yoruba labels, the one shown carries none of the tone marks or subdots of standard Yoruba orthography.

### 7.2 Datasheets

Each language comes with a datasheet (Gebru et al., 2021) covering collection conditions, speaker information where the source documents it, domain, intended and out-of-scope uses, and known limitations, including speaker diversity, domain mismatch, and pseudo-label quality.

### 7.3 Release

The corpus will be deposited in a public repository that assigns a persistent identifier and mirrored on HuggingFace. Two layers are distinguished. The source recordings remain PD-US (public domain in the US), and the `license` field of each record gives that source status. The compiled dataset, that is, the segmentation, filtering, labels, metadata, and packaging, is our own work and is released under CC BY 4.0, so its only condition is attribution. The fine-tuned Whisper-small checkpoints will be released with it. The pipeline code, from download through segmentation, pseudo-labeling, language identification, fine-tuning, and evaluation, is available under the MIT license (a software license, separate from the status of the data) at `https://github.com/arktos-ai-lab/afribase-speech-pipeline`. It applies to any instruction recording with one dominant instruction language and public domain or permissively licensed content.

## 8 Discussion

### 8.1 What AfriBase-Speech enables

*ASR fine-tuning.* Section 6 tests this use directly. Practitioners who need African language training audio without license negotiation now have a documented source, with the caveat that its labels are only as good as the model that produced them.

*TTS bootstrapping.* Slow, clearly enunciated, single-speaker lesson audio is close to studio recording conditions for speech synthesis. Paired with transcripts, it offers a route to open neural voice-synthesis systems for languages that have none. We do not demonstrate this here.

*Phonetic reference.* For field linguists building phonological resources, carefully enunciated lesson speech is a convenient reference for forced alignment tools.

### 8.2 What it cannot do

AfriBase-Speech does not replace natural speech data. It covers at most three speakers per language, a narrow domain, and a teaching register that may not transfer to conversation.

Its language coverage is also narrow: 22 African languages out of more than 2,000. The selection reflects where the US government invested in language training, which followed Cold War and later strategic interests. West and East Africa are well represented; Central and Southern Africa are not. NCHLT partly fills the gap for South Africa.

### 8.3 The license audit as a call to action

The four washing cases are a solvable transparency problem. Platforms display license tags prominently, while EULA and data-card restrictions take manual navigation to find. Repositories could expose restrictions such as TTS bans as structured fields beside the license tag. Until they do, reading the actual terms is the only reliable compliance practice, and we recommend it as standard.

## 9 Conclusion

AfriBase-Speech makes three contributions. It documents license washing in four of 20 screened African speech datasets and separates those cases from two intentional community-protective licenses. It packages US government lesson audio for 22 African languages and Haitian Creole as an NLP-ready dataset of public domain recordings, released under CC BY 4.0, and shows that English instructor speech makes up a substantial share of such audio that must be removed. It evaluates the corpora by fine-tuning Whisper-small for five languages: WER falls for Lingala, Swahili, and Yoruba and rises for Amharic, where the labeling model itself cannot transcribe the language. In the US, works of the federal government need no license and carry no hidden terms.

The broader argument is to build on the public domain tier first, before negotiating CC licenses, auditing EULAs, or trusting what repository metadata implies. Clean provenance enables clean replication.

## Limitations

**Error rates remain very high.** Six of the 15 WER values in Table 3, all for Amharic and Yoruba, exceed 100%. None of these systems is usable for transcription. The contribution is a controlled comparison on a fixed test set, not a deployable recogniser.

**Labels are machine-generated.** Training labels are Whisper large-v3 output, not human transcriptions, and no human checked them. Where large-v3 is weak, as for Amharic, the labels are largely wrong.

**The evaluation is small and single-run.** Five of the 23 inventoried languages are evaluated, each condition once, with no seeds, variance, or confidence intervals. Differences of a few points should not be read as meaningful.

**Language identification is imperfect.** English removal relies on Whisper's language identifier, which we did not validate against human judgement. Some English may remain, and some target-language segments may have been removed.

**Domain mismatch is uncontrolled.** Training audio is slow pedagogical speech; FLEURS is read news. Without an in-domain test set we cannot separate the effect of fine-tuning from that of the mismatch.

**Ewe is packaged but not evaluated.** Its WAXAL split (CC BY 4.0) is the intended held-out set.

**The rights status has a boundary.** Section 105 covers federal employees' work within their duties. Material by contractors, grantees, host-country instructors, or other third parties is excluded. We relied on collection-level provenance, not recording-by-recording authorship checks.

## References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A massively-multilingual speech corpus. In *Proc. LREC 2020*, pages 4218-4222.

Jaco Badenhorst and Febe de Wet. 2022. NCHLT auxiliary speech data for ASR technology development in South Africa. *Data in Brief*, 41:107860.

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 *Proc. SLTU 2014*.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. FLEURS: Few-shot learning evaluation of universal representations of speech. In *Proc. IEEE SLT 2022*, pages 798-805.

Abdoulaye Diack, Perry Nelson, Kwaku Agbesi, Angela Nakalembe, Mohamed MohamedKhair, Vusumuzi Dube, Tavonga Siyavora, Subhashini Venugopalan, Jason Hickey, Uche Okonkwo, Abhishek Bapna, Isaac Wiafe, and others. 2026. WAXAL: A large-scale multilingual African language speech corpus. arXiv:2602.02734.

Sanchit Gandhi, Patrick von Platen, and Alexander M. Rush. 2023. Distil-Whisper: Robust knowledge distillation via large-scale pseudo labelling. arXiv:2311.00430.

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

Internet Archive. 2014. Terms of use, privacy policy, and copyright policy (terms of use dated 31 December 2014). https://archive.org/about/terms.php, accessed 30 October 2026.

Ussen Kimanuka, Ciira wa Maina, and Osman Büyük. 2023. Speech recognition datasets for low-resource Congolese languages. *Data in Brief*, 52:109796.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Tajudeen Kolawole, Taiwo Fagbohungbe, and others. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of EMNLP 2020*.

Chijioke I. Okorie and Melissa Omino. 2025. Addressing inequitable openness in licences for sharing African data and datasets through the Nwulite Obodo Open Data Licence. *Law, Technology and Humans*, 7(3):94-110.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. LibriSpeech: An ASR corpus based on public domain audio books. In *Proc. ICASSP 2015*.

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.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proc. ICML 2023*, pages 28492-28518.

Jenalea Rajab, Anuoluwapo Aremu, Everlyn Asiko Chimoto, Dale Dunbar, Graham Morrissey, Fadel Thior, Luandrie Potgieter, Jessico Ojo, Atnafu Lambebo Tonja, Maushami Chetty, Wilhelmina Nekoto, Pelonomi Moiloa, Jade Abbott, Vukosi Marivate, and Benjamin Rosman. 2025. The Esethu framework: Reimagining sustainable dataset governance and curation for low-resource languages. In *Proc. ACL 2025*, pages 30763-30776.

Ernst A. P. van Gassen. 2026a. Open but unvetted: The ethics of African language data. In *Proc. RAIL 2026 at LREC 2026*, pages 128-139. ELRA.

Ernst van Gassen. 2026b. Archival ASR for African broadcasting: Automatic transcription of VOA African-language audio. To appear in *Proc. DHASA 2026* (accepted; submission 140).

## A Tier 1a lesson audio inventory

| Language | Source | Files | Hours | FLEURS |
|---|---|---|---|---|
| Amharic | FSI | 180 | 21.2* | am_et |
| Bambara | Peace Corps | 13 | 1 | n/a |
| Chewa | Peace Corps | 17 | 1 | ny_mw |
| Ewe | Peace Corps | 104 | 10 | n/a |
| Haitian Creole | DLI | 59 | 3 to 5 | n/a |
| Hausa | FSI | 10 | 3.6* | ha_ng |
| Igbo | FSI | 64 | 4 to 6 | ig_ng |
| Jola-Fonyi | Peace Corps | 22 | 1.5 | n/a |
| Krio | FSI, Peace Corps | 88 | 4 to 6 | n/a |
| Lingala | FSI | 24 | 9.5* | ln_cd |
| Luganda | Peace Corps | 12 | 1 | lg_ug |
| Malagasy | Peace Corps | 15 | 1 | n/a |
| Mandinka | Peace Corps | 25 | 2 | n/a |
| Mooré | Peace Corps | 21 | 1.5 | n/a |
| Pulaar | Peace Corps | 31 | 2 | n/a |
| Setswana | Peace Corps | 43 | 3 | n/a |
| SiSwati | Peace Corps | 9 | 0.75 | n/a |
| Soninke | Peace Corps | 20 | 1.5 | n/a |
| Swahili | FSI | 132 | 15.1* | sw_ke |
| Twi | Peace Corps | 19 | 1.5 | n/a |
| Wolof | Peace Corps | 27 | 2 | wo_sn |
| Yoruba | FSI | 279 | 15.6* | yo_ng |
| Zarma | Peace Corps | 21 | 1.5 | n/a |

Table 4: Tier 1a public domain lesson audio, in alphabetical order. *Measured on the downloaded audio of the five evaluation languages, which come from the FSI courses only; other hours are estimates from file metadata, except Ewe, which was measured.

## B Datasets screened in the license audit

| Dataset | Platform tag | Finding |
|---|---|---|
| *License washing (Section 3)* | | |
| DDD-Kenya | CC BY 4.0 | EULA bans commercial AI, voice cloning, biometrics |
| Swivuriso (za-african-next-voices) | CC BY 4.0 | gated EULA bans TTS and voice cloning |
| West African Radio Corpus (SLR105) | CC BY-SA 4.0 | third-party broadcast content the depositors cannot license |
| Gamayun Kenyan Swahili | custom, "open" | non-commercial only |
| *Intentional community-protective license* | | |
| Esethu / isiXhosa (Lelapa AI) | Esethu | benefit sharing for OECD commercial users |
| DhoNam Dholuo | NOODL | royalty or benefit sharing for OECD commercial users |
| *Restrictive, and tagged as such* | | |
| Zambezi Voice | CC BY-NC-ND 4.0 | non-commercial, no derivatives |
| NaijaVoices | CC BY-NC-SA 4.0 | non-commercial |
| AfriSpeech-200 | CC BY-NC-SA 4.0 | non-commercial; African-accented English, out of scope |
| *Tag consistent with terms* | | |
| NCHLT Auxiliary (Aux1, Aux2) | CC BY 4.0 | consistent |
| NCHLT Hansard | government open data | consistent |
| Lwazi ASR (telephone) | CC BY 3.0 / 2.5 ZA | consistent |
| Lwazi III TTS | CC BY 4.0 / 2.5 ZA | consistent |
| ALFFA (SLR25) | MIT | consistent |
| Fongbe speech | CC0 | consistent |
| Kallaama (SLR151) | CC BY 4.0 | consistent |
| Luganda/Swahili TTS subset (Mendeley) | CC BY 4.0 | consistent |
| Common Voice (African languages) | CC0 | consistent, per clip |
| *Other* | | |
| WAXAL | CC BY 4.0 (paper) | per-partner split: CC BY 4.0 for Ghana, CC BY-SA 4.0 elsewhere (van Gassen, 2026a) |
| Thiomi | CC BY 4.0 (paper) | not yet released; terms unverifiable |

Table 5: The 20 African speech datasets screened, grouped by finding. Terms checked between June and October 2026. NC: non-commercial; ND: no derivatives; SA: share-alike; ZA: South Africa.