Archival ASR for African Broadcasting: Automatic Transcription of VOA African-Language Audio

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

Abstract

Voice of America (VOA) broadcast in African languages for more than six decades through dedicated language services: newsrooms that each produced radio, television, and web content in one language. Its own output is in the public domain under United States (US) law, because works of US federal government employees are not protected by US copyright. The March 2025 shutdown of VOA's Africa division has frozen the archive. We crawl all 16 African-language services (Afaan Oromo, Amharic, Bambara, French, Fulfulde, Hausa, Kinyarwanda and Kirundi, Lingala, Ndebele, Portuguese, Sango, Shona, Somali, Swahili, Tigrinya, and Wolof), plus Haitian Creole. We transcribe the eight languages covered by OpenAI's Whisper speech recognition model into 364,815 utterances; Hausa (220.4 hours) and Portuguese (235.7 hours) are the largest. Meta's Massively Multilingual Speech (MMS) model covers every service language except Ndebele, and we add MMS transcripts of 3,000-segment samples for Afaan Oromo, Tigrinya, Kinyarwanda, and Bambara. Audio on modern VOA pages loads through JavaScript and resists standard download tools; a legacy content management system recovers most of what we archive. On FLEURS (Few-shot Learning Evaluation of Universal Representations of Speech), a read-speech benchmark, MMS has a lower word error rate (WER) than Whisper for all six languages where both run, 14.0 to 44.3% against 35.2 to 157.9% for Whisper large-v3, and reaches 38.9% on Wolof and 63.9% on Afaan Oromo. Fine-tuning Whisper-small on public domain lesson audio lowers its WER for two of four languages but does not reach large-v3. On the archive itself, transcript quality tracks each language's coverage in the base model. We release the audio, provenance metadata, transcripts with per-language quality grades, and the pipeline code.

Keywords: speech recognition, African languages, archival preservation, Voice of America, public domain, Whisper, MMS

1 Introduction

The Voice of America (VOA) was established in 1942 and at its peak broadcast in 49 languages (USAGM, n.d.). VOA organised this output by *language service*: a newsroom that produces radio, television, and web content in one language for one audience. By August 2024 there were 16 African-language services. They covered Afaan Oromo, Amharic, Bambara, French (for Africa), Fulfulde, Hausa, Kinyarwanda and Kirundi (one shared service), Lingala, Ndebele, Portuguese (for Africa), Sango, Shona, Somali, Swahili, Tigrinya, and Wolof. This paper studies the audio and video recordings and the article pages that these services published on VOA's websites, which we recovered by crawling. VOA staff are United States (US) federal government employees. Under 17 U.S.C. § 105, works that US federal government employees make as part of their duties are not protected by US copyright, so they are in the public domain: anyone may copy and reuse them. For VOA's own material it is a stronger and clearer basis for reuse than the Creative Commons (CC) licence labels that third-party aggregators sometimes attach to VOA content (see van Gassen, 2026a).

In March 2025 the Africa division went dark, and nobody now maintains the archive that remains online. Despite its clear status under US law it has seen little use as a speech resource: no transcription exists, the audio is scattered across thousands of article pages, and automatic speech recognition (ASR) quality varies widely by language.

We make four contributions. First, we assemble the archive: we crawl all 16 African-language services plus Haitian Creole, including the pages of a legacy content management system (CMS) that served VOA articles from about 2005 to 2016, and record per-file provenance. Second, we charac-

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.

terise it acoustically by signal-to-noise ratio (SNR), speech fraction, and duration, which sets realistic expectations for transcription. Third, we transcribe the eight languages covered by OpenAI's Whisper speech recognition model, add samples from Meta's Massively Multilingual Speech (MMS) model for four languages Whisper does not cover, and compare four model conditions on FLEURS, the Few-shot Learning Evaluation of Universal Representations of Speech benchmark, which provides human-verified transcriptions of read news sentences. One condition is Whisper-small fine-tuned on AfriBase-Speech (van Gassen, 2026b), a corpus of public domain US government lesson audio for African languages, accepted at the Digital Humanities Association of Southern Africa (DHASA) 2026 conference. Fourth, we release audio, metadata, transcripts, and pipeline code.

We report error rates as word error rate (WER): the number of word substitutions, deletions, and insertions needed to turn the system output into the reference, divided by the number of reference words. Because insertions are counted, WER can exceed 100% when a system produces more wrong words than the reference contains.

Our framing is preservation: imperfect automatic transcription of a closed archive is more useful than none, provided its failure modes are documented per language.

The rest of the paper is organised as follows. Section 2 reviews related work. Section 3 describes how the archive was assembled and segmented and gives its statistics. Section 4 presents the transcription conditions, the FLEURS results, and the quality of the archive transcripts. Section 5 describes the state of the archive after the 2025 shutdown, Section 6 the release, and Section 7 the conclusions, followed by the limitations. Appendix A gives early acoustic statistics, and Appendix B gives script-purity figures and example transcripts.

2 Background and Related Work

2.1 VOA as a language resource

VOA has mainly been used in natural language processing (NLP) as a *text* resource. The Multilingual Open Text project collected over 2.8 million VOA news articles in 44 languages (Palen-Michel et al., 2022), released under the Creative Commons Attribution 4.0 licence (CC BY 4.0), which permits reuse if the author is credited, although the source material is public domain. VOA's terms of use state that "all text, audio and video material produced exclusively by the Voice of America is in the public domain," and exclude licensed wire and third-party material.[1] We know of no prior systematic ASR study of VOA African-language audio.

2.2 Radio and broadcast archives for ASR

Broadcast archives are an established source of speech data where read or crowdsourced speech is scarce. RadioTalk transcribed US talk radio at scale with an off-the-shelf recogniser (Beeferman et al., 2019). For African languages, Saeb et al. (2017) built keyword-spotting systems over Ugandan community radio for humanitarian monitoring with very little transcribed data. Doumbouya et al. (2021) used West African radio archives to pretrain speech models for Maninka, Pular, and Susu. Kimanuka et al. (2023) released 741 hours of unlabelled Congolese radio; pretraining on it lowered Lingala 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). Earlier resources were also built for languages we study: Gauthier et al. (2016) collected data and built recognisers for Swahili, Hausa, Amharic, and Wolof, from collected speech rather than broadcast archives. Our work differs in its source and its goal. VOA audio is public domain under US law, so it can be redistributed as well as used there. We transcribe the archive for preservation and search rather than to train a single recogniser.

2.3 Multilingual ASR for African languages

Whisper (Radford et al., 2023) covers 99 languages, among them Amharic, Hausa, Lingala, Shona, Somali, Swahili, and Yoruba, but its training data for these is thin. MMS extends recognition to over 1,100 languages with a shared model and small per-language adapters (Pratap et al., 2024); its training data include the FLEURS training splits. Read-speech benchmarks such as FLEURS (Conneau et al., 2023), with about 12 hours per language, and crowdsourced corpora such as Common Voice (Ardila et al., 2020) are the usual evaluation and training sets. Earlier corpus work for individual languages, such as the Amharic read-speech corpus of Abate et al. (2005), predates these models. Koe-

[1] https://www.voanews.com/terms-use-and-privacy-notice, accessed 17 October 2026.

Service	ISO	FLEURS	Whisper	MMS
Lingala	lin	ln_cd	yes	yes
Hausa	hau	ha_ng	yes	yes
Swahili	swh	sw_ke	yes	yes
Amharic	amh	am_et	yes	yes
Somali	som	so_so	yes	yes
Shona	sna	sn_zw	yes	yes
Haitian Creole	hat	n/a	yes	yes
Portuguese (Africa)	por	n/a	yes	yes
French (Africa)	fra	n/a	yes	yes
Wolof*	wol	wo_sn	no	yes
Afaan Oromo	orm	om_et	no	yes
Ndebele	nde	n/a	no	no
Tigrinya	tir	n/a	no	yes
Kinyarwanda/Kirundi	kin	n/a	no	yes
Bambara	bam	n/a	no	yes
Sango*	sag	n/a	no	yes
Fulfulde*	ful	n/a	no	yes

Table 1: The 17 crawled services. "FLEURS" gives the test-set code where one exists; "Whisper" and "MMS" mark language coverage in Whisper and in MMS (`mms-1b-all` adapters). Rows run from the Whisper-covered languages to the rest. *Published inside the French-to-Africa site, `voaafrique.com`, without a dedicated domain.

necke et al. (2024) show that Whisper sometimes hallucinates text with no basis in the audio; we observe the same failure on the archive (Section 4.3). More recently, Nahabwe et al. (2026) fine-tuned Whisper, XLS-R, MMS, and W2v-BERT on 1 to 400 hours of data for 13 African languages; they report that MMS and W2v-BERT are more data-efficient at very low resource levels and Whisper performs best at mid-resource levels. Our zero-shot FLEURS comparison fits MMS's advantage where little language-specific data exists, but we do not fine-tune MMS. van Gassen (2026b), accepted at this conference, package public domain US government lesson audio for African languages. We use the fine-tuned models from that work as one evaluation condition.

3 Archive Assembly

The pipeline (Figure 1) starts from the web addresses (uniform resource locators, URLs) and numeric identifiers (IDs) of VOA articles. Table 1 lists the 17 services we crawl, with their ISO 639-3 language codes (three-letter codes of the International Organization for Standardization), and whether Whisper and MMS cover each language. Whisper covers nine. MMS covers all except Ndebele, which leaves Ndebele as the one service with audio that no model we tested can transcribe.

3.1 Finding the audio

Each service publishes a `sitemap.xml` index that lists article URLs. We collect all article URLs from it, sort them newest first, and attempt to download embedded media from each page, first through a direct media link in the page source and otherwise with the `yt-dlp` downloader. Article text is saved whether or not audio is found.

On modern VOA pages (article IDs above roughly six million, from about 2018), audio is inserted by JavaScript after the page loads. The hypertext markup language (HTML) source that a crawler receives contains only an empty audio element with `src="about:blank"`, and `yt-dlp` then returns a silent placeholder file of 100.1 seconds. A quality gate rejects any file quieter than -40 decibels relative to full scale (dBFS) or lasting 100.1 ± 0.5 seconds. After this gate, modern article pages yield no usable audio for any service we tested. The Swahili service embeds a live radio stream instead of per-article clips, which yields nothing archivable either.

The audio that we do recover comes mostly from `editorials.voa.gov`, the legacy CMS, which serves pages from roughly 2005 to 2016 with plain links to MP3 audio and MP4 video files. We reach it by scanning article IDs in sequence. This ID space is shared by every VOA service worldwide, including Arabic, Urdu, and Persian. Each page therefore passes a language gate: its HTML `lang` attribute must match the target service, and a page whose language cannot be verified is rejected. Combining the legacy scan with sitemap enumeration, capped at 4,000 candidate articles per service, recovers substantially more audio than either source alone.

3.2 Segmentation and archive statistics

Downloaded files are split into utterance-length segments at pauses. Short article clips are split at silences of at least 600 milliseconds below -40 dBFS, and a segment is kept if it lasts 1 to 30 seconds, has an SNR of at least 6 decibels (dB), and is mostly speech. SNR is estimated from frame energy, comparing the loudest quarter of 20-millisecond frames with the quietest quarter. Full broadcast recordings rarely pause for that long, so for them we split at silences of 200 milliseconds below -38 dBFS and drop the SNR floor; continuous speech gives a low SNR estimate even when the recording is clean (Appendix A).

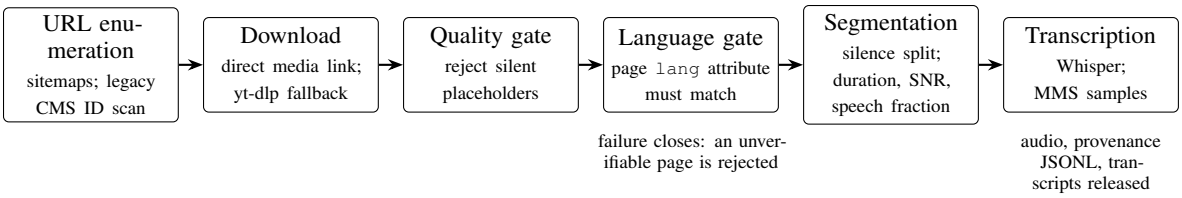

Figure 1: Archive pipeline, from VOA article URLs to released transcripts. Each stage records its decision per file in the provenance metadata.

Hausa, Portuguese, and Haitian Creole dominate (Table 2). The legacy CMS ID range around 3.8 to 4.5 million is rich in Hausa content, and the Portuguese and Haitian Creole services are large. Swahili, Amharic, Somali, Shona, and Lingala yield few segments despite thousands of files, because most of their files are long continuous broadcasts. Such recordings contain few clean pauses and little silence, so few stretches pass the segment filters. Appendix A gives per-language medians from an early characterisation pass. French recovered no new segmentable audio, because its candidate pages either duplicated the Wolof, Sango, and Fulfulde crawls of the same domain or fell in the JavaScript range.

For Afaan Oromo, Tigrinya, Kinyarwanda/Kirundi, and Bambara, the full crawl saved 2,701, 2,374, 2,076, and 3,640 article pages. A second crawl, capped at 120 media files per service, found downloadable audio on 40 to 59% of the pages it probed (120 files from 203 to 303 pages), and we segmented those files with the broadcast settings.

Lang.	Files	Segments	Hours	Status
lin	25	751	0.96	transcribed
hau	1,319	191,335	220.36	transcribed[a]
swh	1,445	134	1.35	transcribed
amh	3,675	122	1.27	transcribed
som	2,956	24	0.24	transcribed
sna	3,139	29	0.25	transcribed
hat	1,900	55,680	122.36	transcribed
por	2,706	116,773	235.73	transcribed
fra	0	0	0	no new audio
wol	1,677[b]	n/a	n/a	archived
orm	120[c]	19,465	17.26	MMS sample
nde	2,763[b]	663,996	n/a	archived
tir	120[c]	26,746	21.09	MMS sample
kin	120[c]	48,509	40.88	MMS sample
bam	120[c]	15,046	14.22	MMS sample
sag	0	n/a	n/a	text only
ful	0	n/a	n/a	text only

Table 2: Archive statistics, in the order of Table 1. "Files" counts downloaded audio and video files. Segments and hours are given where segmentation was run; Ndebele was segmented but cannot be transcribed. "Text only": all candidate pages carried JavaScript-loaded audio. [b]Article pages saved by the full crawl; the share carrying audio was not measured. [c]Second crawl, capped at 120 media files per service. [a]33 segments from three misattributed source files are withheld (Section 4.3).

4 Transcription and Evaluation

4.1 Conditions

No human transcriptions of VOA audio exist, so we cannot measure WER on the archive directly. We measure it on the FLEURS test split (Conneau et al., 2023), which provides human-verified transcriptions of read news sentences, and treat the result as a proxy. Four of the transcribed languages (Lingala, Hausa, Swahili, Amharic) have both a FLEURS test set and fine-tuned models from the companion paper. Somali and Shona have FLEURS test sets but no fine-tuned model. FLEURS has no Haitian Creole or African Portuguese test set. Wolof and Afaan Oromo have FLEURS test sets but no Whisper coverage.

Small zero-shot is Whisper-small (244 million parameters) applied without fine-tuning. **Small fine-tuned** is Whisper-small fine-tuned for 10 epochs on AfriBase-Speech lesson audio (van Gassen, 2026b). These checkpoints were trained on lesson audio segmented at pauses, with Whisper large-v3 transcriptions as training labels. The companion paper's final models additionally remove English instructor speech before training; the archive transcripts here predate that step. **Large-v3 zero-shot** is Whisper large-v3 (1.55 billion parameters) without fine-tuning. Every Whisper condition decodes with the target language set in advance rather than detected automatically. **MMS** is `mms-1b-all` (1 billion parameters), a connectionist temporal classification model, with the target language's adapter loaded and greedy decoding without a language model. All four are scored with the same text normalisation (lower-casing, punctuation removal, whitespace collapsing). For transcribing the archive, we used the fine-tuned Whis-

	Whisper-small		Large-v3	MMS
Language	**zero-shot**	**fine-tuned**	**zero-shot**	**1b-all**
Lingala	89.9	78.2	72.0	14.7
Hausa	90.6	91.2	86.2	25.1
Swahili	73.2	52.1	35.2	14.0
Amharic	115.7	166.9	157.9	28.0
Somali	n/a	n/a	89.1	44.3
Shona	n/a	n/a	114.6	22.7
Wolof	n/a	n/a	n/a	38.9
Afaan Oromo	n/a	n/a	n/a	63.9

Table 3: WER (%) on the FLEURS test split, in the order of Table 1. Lower is better; values above 100% mean the output contains more errors than the reference has words. Somali and Shona have no fine-tuned model; Wolof and Afaan Oromo are not covered by Whisper. The Afaan Oromo test split has only 41 utterances.

Lang.	**Model**	n	**Match**	**Distinct**
lin	small-ft	751	20.9	72.4
hau	small-ft	191,302	27.4	95.3
swh	small-ft	134	100.0	5.2
amh	small-ft	122	0.0[a]	98.4
som	large-v3	24	0.0	8.3
sna	large-v3	29	0.0	69.0
hat	large-v3	55,680	74.8	86.9
por	large-v3	116,773	97.1	77.3
orm	MMS	3,000	88.1	95.3
tir	MMS	3,000	83.5	98.3
kin	MMS	3,000	74.6[b]	99.2
bam	MMS	3,000	75.1	99.6

Table 4: Reference-free quality of the archive transcripts (%). "Match": GlotLID assigns the transcript to the target language. "Distinct": share of a language's transcripts that are unique. "small-ft": Whisper-small fine-tuned on lesson audio. [a]GlotLID returns undetermined Ethiopic script for all 122. [b]A further 10.5% are identified as Kirundi, which the service also broadcast.

per model where one existed and large-v3 otherwise; for Afaan Oromo, Tigrinya, Kinyarwanda, and Bambara we transcribe a random sample of 3,000 segments per language (2.4 to 2.9 hours) with MMS.

4.2 Results

Fine-tuning Whisper-small on lesson audio lowers its WER for Lingala (89.9% to 78.2%) and Swahili (73.2% to 52.1%), leaves Hausa essentially unchanged, and raises Amharic sharply (Table 3). Against the far larger zero-shot large-v3, the fine-tuned small model is worse for every language. The gap is widest for Swahili, where large-v3 is already comparatively strong (35.2%). A few hours of lesson audio narrows the distance between the small and the large model but does not close it.

MMS has the lowest WER of the four systems for every language with a Whisper result, by a wide margin: 14.7% against 72.0% for large-v3 on Lingala, 25.1% against 86.2% on Hausa, and 28.0% against 157.9% on Amharic. Part of this advantage is one of domain, since MMS was trained on the FLEURS training splits and the Whisper models were not. A gap this large is unlikely to be domain alone: for Amharic and Shona, Whisper's WER exceeds 100%, while MMS stays below 30%. For a practitioner choosing a model for this archive, MMS is the better default for every service language it covers.

Among the Whisper conditions, Amharic is the exception in both directions. Its large-v3 WER is above 100% even before fine-tuning, so the training labels that large-v3 produced for the fine-tuned model are themselves mostly wrong. The small model learns to emit Ge'ez syllable sequences that look plausible but do not match the audio. Decoding mode does not explain the Amharic result. With automatic language detection, large-v3 scores 140.7% instead of 157.9% on FLEURS, but on the lesson audio it writes 98.5% of Amharic labels in Arabic rather than Ge'ez script (Appendix B).

4.3 Transcript quality on the archive

FLEURS is read speech, so we also measure the transcripts that the pipeline produced for the archive itself, without references (Table 4). For each transcript we record whether the GlotLID text language identifier (Kargaran et al., 2023) assigns it to the target language, and for each language the share of distinct transcripts, which falls when a model repeats a fixed output. A language match shows that the output is text in the right language, not that it matches the audio. It is a weaker signal for MMS, whose adapters can only emit characters of the target language.

Quality follows the base model's coverage of each language, not the fine-tuning condition. The fine-tuned Whisper-small transcripts of Hausa and Lingala are in the target language only 27.4 and 20.9% of the time, and often consist of strings with plausible sound patterns that are not real words. Swahili always matches, but only 5.2% of its transcripts are distinct: the model falls into clause-level repetition loops over music or station identification, despite a decoding constraint against repeated three-word sequences. Among the languages transcribed with large-v3, Haitian

Creole and Portuguese produce fluent, on-topic text. Somali and Shona do not. Every Somali segment reads `Altyazı M.K.` (Turkish for "Subtitles M.K."), and the Shona output consists of video outros such as `Thank you for watching!` and `Shabbat shalom.`, or of long runs of characters from unrelated scripts. These are artefacts of Whisper's training data, the hallucination pattern described by Koenecke et al. (2024). Both samples are small (24 and 29 segments). The MMS samples match the target language in 74.6 to 88.1% of cases, and 95.3 to 99.6% of them are distinct. Appendix B reproduces examples.

The language gate works at page level, and the shared legacy ID space lets pages from other services through occasionally. A spot check of the earliest Hausa segments found three source files whose audio large-v3 transcribes as fluent Turkish. We withhold their 33 segments from the release.

We grade the released transcripts accordingly. Haitian Creole and Portuguese are usable as approximate search text. Hausa, Lingala, Swahili, and Amharic are noisy indices for exploration. Somali and Shona carry no information about the audio and are released only so that the record is complete. The MMS samples for Afaan Oromo, Tigrinya, Kinyarwanda, and Bambara are approximate text whose accuracy is unmeasured except for Afaan Oromo, where the FLEURS proxy is 63.9% WER.

Wolof, Sango, and Fulfulde were published inside the French-to-Africa site, whose pages all declare French as their language. Text language identification found no Wolof in a random sample of 197 article pages from the site, so we cannot confirm that Wolof material survives in the online archive, and we do not transcribe it.

5 The VOA Archive as a Preservation Resource

On 14 March 2025, Executive Order 14238 directed that the non-statutory functions of seven agencies, among them VOA's parent agency, the US Agency for Global Media, be reduced to the minimum the law requires (Executive Order 14238, 2025). VOA's Africa division stopped broadcasting and web publishing shortly afterwards, including the Fulfulde service launched only seven months earlier. The archive is now closed: no new content will be produced, and what remains online has no funded team behind it.

First, content is deleted: even while VOA operated, sitemap entries often returned errors within days of publication, and nobody now remains to notice or contest a removal. Second, domains may lapse, since dormant government web infrastructure has no assured maintenance. Third, article text, the only written record accompanying the audio, disappears with the pages. Our pipeline therefore captures audio and article metadata in one pass, and it can be rerun on any later digitisation of the pre-2005 archive, which reaches back decades further than what the web exposes. With no new content to track, systematic capture is a one-time rescue, which makes completing it more urgent.

6 Release

We release per-language audio segments with provenance metadata in JSON Lines (JSONL) format (source URL, article ID, verified page language, segment timing, SNR), the transcripts with the quality grades of Section 4.3 (each record names the model that produced it), and the crawl, segmentation, transcription, and evaluation code. Two layers are distinguished. The source audio is in the public domain under US law (17 U.S.C. § 105). The compiled dataset, that is, the segmentation, metadata, transcripts, and quality grades, is our own work and is released under the Creative Commons Attribution 4.0 licence (CC BY 4.0), so its only condition is attribution. The audio and transcripts will be deposited on HuggingFace, a hosting platform for datasets and models, alongside the AfriBase-Speech corpus. The code is released separately under the MIT licence, a software licence, at `https://github.com/arktos-ai-lab/afribase-speech-pipeline`.

7 Conclusion

We crawled all 16 VOA African-language services plus Haitian Creole and transcribed the eight Whisper-covered languages, producing 364,815 utterances, and added MMS samples for four services Whisper does not cover. Hausa and Portuguese account for most of the volume, and the legacy CMS, not the JavaScript-driven modern pages, is what makes the archive recoverable. On FLEURS, MMS outperforms every Whisper condition for all six languages where both run. Fine-tuning Whisper-small on public domain lesson audio improves it for two of four languages but leaves it behind zero-shot

large-v3 for all four. On the archive, the usefulness of a transcript depends on the base model's coverage of the language. Two steps follow: retranscribing the Whisper-covered services with MMS, and evaluating directly on a human-transcribed sample of VOA audio. The VOA Africa archive will not grow again, which makes capturing what is still online a time-bounded task for the African NLP community.

Limitations

Transcript quality is low for six of eight languages. Only Haitian Creole and Portuguese produce fluent output. The Hausa, Lingala, Swahili, and Amharic transcripts are noisy, and the Somali and Shona transcripts are training-data artefacts. These transcripts were produced with Whisper and have not been regenerated with MMS, despite its lower FLEURS error rates.

MMS coverage is a sample. For Afaan Oromo, Tigrinya, Kinyarwanda, and Bambara we transcribe 3,000 segments each from a crawl capped at 120 files, not the full service. The `mms-1b-all` model weights are released under the Creative Commons Attribution-NonCommercial 4.0 licence (CC BY-NC 4.0), which does not permit commercial use; this is a licence on the model files, not on the audio. Whether it extends to the transcripts the model produced is uncertain, and we do not decide it. Each transcript record names the model.

WER is measured on a proxy. FLEURS is read speech in a different domain from VOA broadcasts. We have no human-transcribed VOA sample, so no WER figure here describes the released transcripts directly.

The fine-tuning comparison is confounded by model size. The fine-tuned model has about one sixth of large-v3's parameters. Table 3 includes zero-shot Whisper-small to separate the two effects, but each condition was run once, without seeds or confidence intervals.

Language identity is verified per page, not per segment. We did not run audio language identification over the release; the GlotLID grades of Table 4 check the transcript text only. Contamination beyond the three Hausa files we found is possible and unquantified.

Coverage is partial. Haitian Creole and Portuguese have no FLEURS test set, so their transcripts carry no WER estimate, and the statistics in Appendix A predate the final crawl. Counts are a snapshot of a source from which pages keep disappearing.

Public domain status has a boundary. 17 U.S.C. § 105 covers works of US federal employees made within their duties. Freelance and stringer contributions, licensed wire copy, and music inside a broadcast fall outside that rule, and their status is not established here. We relied on the collection-level position and did not clear rights segment by segment.

References

Solomon Teferra Abate, Wolfgang Menzel, and Bairu Tafila. 2005. An Amharic speech corpus for large vocabulary continuous speech recognition. In *Proc. Interspeech 2005*, pages 1601-1604.

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.

Doug Beeferman, William Brannon, and Deb Roy. 2019. RadioTalk: A large-scale corpus of talk radio transcripts. In *Proc. Interspeech 2019*, pages 564-568.

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.

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 *Proc. AAAI 2021*, 35(17):14757-14765.

Executive Order 14238. 2025. Continuing the reduction of the federal bureaucracy. 90 Fed. Reg. 13043, 19 March 2025.

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. In *Proc. LREC 2016*, pages 3863-3867.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. 2023. GlotLID: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155-6218.

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

Allison Koenecke, Anna Seo Gyeong Choi, Katelyn X. Mei, Hilke Schellmann, and Mona Sloane. 2024. Careless Whisper: Speech-to-text hallucination harms. In *Proc. ACM FAccT 2024*. doi:10.1145/3630106.3658996.

Alvin Nahabwe, Sulaiman Kagumire, Denis Musinguzi, Bruno Beijuka, Jonah Kyagaba, Peter Nabende, Andrew Katumba, and Joyce Nakatumba-Nabende. 2026. Benchmarking automatic speech recognition models for African languages. In *Proceedings of Machine Learning Research*, volume 302 (DLI 2025 Research Track), pages 1-19. https://proceedings.mlr.press/v302/nahabwe26a.html.

Chester Palen-Michel, June Kim, and Constantine Lignos. 2022. Multilingual Open Text release 1: Public domain news in 44 languages. In *Proc. LREC 2022*, pages 2080-2089.

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.

Armin Saeb, Raghav Menon, Hugh Cameron, William Kibira, John Quinn, and Thomas Niesler. 2017. Very low resource radio browsing for agile developmental and humanitarian monitoring. In *Proc. Interspeech 2017*, pages 2118-2122.

USAGM. n.d. Voice of America. U.S. Agency for Global Media. https://www.usagm.gov/networks/voa/.

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. AfriBase-Speech: Auditing, tiering, and packaging African speech resources from government public domain archives. In *Proc. Digital Humanities Association of Southern Africa (DHASA) 2026*. Accepted.

A Early acoustic characterisation

Table 5 reports per-language medians from an early pass over raw downloaded files, before the full crawl. Lingala statistics come from short clips on the legacy CMS; the other languages are full broadcast recordings of 15 to 60 minutes. Low SNR for Swahili, Amharic, Somali, and Shona reflects continuous broadcast audio with almost no silence (speech fraction above 0.97), not noise: the SNR estimate compares the loudest and quietest quarters of frames, and is low when speech never pauses. Lingala's median segment of 3.0 seconds reflects the sentence-level pauses of VOA news delivery.

In a 200-article pilot that preceded the full crawl, Somali pages embedded extractable audio at close to 100%, Wolof at about 18% before language filtering, Hausa at about 15%, and Lingala at about 8.5%. Amharic, Shona, and Afaan Oromo showed none in that sample but yielded thousands of files in the full crawl. Swahili's live-stream embedding, noted in Section 3, helps explain why its 1,445 files yield only 134 segments.

Lang.	SNR (dB)	Speech frac.	Duration (s)
lin	20.3	0.639	3.0
hau	9.4	0.956	602.6
swh	3.6	0.974	1798.4
amh	2.6	0.996	3598.0
som	3.6	0.974	1799.2
sna	2.6	0.996	1798.0

Table 5: Median acoustic statistics per language from the early pass (raw files), in the order of Table 1.

B Script purity and example transcripts

Script purity. With automatic language detection, Whisper large-v3 writes 98.5% of Amharic training labels in Arabic script rather than Ge'ez. Setting the target language in advance raises the share of labels in the correct script as follows: Amharic 1.5% to 95.0%, Hausa 68.1% to 98.1%, Swahili 95.3% to 99.9%, Lingala 97.3% to 99.9%. Correct script does not imply correct words.

Recurring errors in Lingala and Hausa. Names of politicians, countries, and cities are transcribed inconsistently; regional political figures often come out as a French or English approximation. Broadcasts switch to English or French for technical terms and organisation names (United Nations, International Monetary Fund, World Health Organization), which the model sometimes renders in a third language. Station identifications such as "Hii ni sauti ya Amerika" and programme titles are frequent in the archive but rare in FLEURS, and lesson audio cannot teach them.

Examples. Fine-tuned Whisper-small, Hausa: "Tara danyu, rama te gara bababa." Plausible Hausa sound patterns, but not words. Fine-tuned Whisper-small, Swahili: "Kwa hivyo, kwa hiyo, kama kwa kwa, kuma kwa.", repeated almost verbatim

across dozens of consecutive segments. Large-v3, Haitian Creole: “Mat sa Hamas vle, ak sa yo aksepte. Ak sa Izrayel vle ak sa li aksepte.” Large-v3, Portuguese: “E esta é a primeira vez que um presidente americano pisa em solo angolano.”, from coverage of a US presidential visit to Angola. Large-v3, Somali: all 24 segments read `Altyazı M.K.` Large-v3, Shona: `Thank you for watching!` and `Shabbat shalom.` (5 segments each), and several long runs of Malayalam, Gujarati, Tibetan, and Thai characters. The fine-tuned models attempt the language and fail; for Somali and Shona, large-v3 does not attempt it at all.