JML: A Large-Scale Fongbe Speech Corpus for ASR and Tonal Modeling

Fréjus A. A. Laleye[5] Gaëlle P. Gangbe[2] Imane A. ADEOTI[2]
Izzoudine M. L. Yahaya[2] Fabroni B. Yoclounon[4]
Jean-Eudes D. Awanvoeke[3] Kevin Degila[1] Ferdinand Sodokin[3]
Daouda B. Zakari[3] Anziz A. ADEHAN[2] Mathieu SOGLONOU[3]
Marc-André Loko[2] Mahuna Akplogan[1]
[1]Ministry of Digital Transformation and Innovation, Benin
[2]Agency for Information Systems and Digital Affairs (ASIN), Benin
[3]Regional Digital Lab, Institute for Inclusive Digital Africa (IIDiA), Benin
[4]IamYourClounon, Benin
[5]Opscidia, France

October 30, 2026

Abstract

Fongbe is a tonal Gbe language spoken mainly in Benin, and it remains poorly served by speech technology. Existing public resources total a few tens of hours, and none was designed for studying tone at scale. We present a large-scale Fongbe speech dataset built for automatic speech recognition and tonal modeling. The corpus contains about 700 hours of speech through a web platform, from 2,420 registered contributors across the twelve departments of Benin, who describe photographs or say written French sentences in Fongbe. About 351 hours are transcribed with tone marks, and 159 validated hours, in the frozen export used here, were checked by a second person. We describe the collection protocol, the orthographic and tonal conventions, and the Unicode normalization applied to transcripts. We report an audit of annotation quality: before review, 25.6% of written word forms occurred with more than one tonal spelling and tone-mark density varied from 0.377 to 0.495 marks per vowel across annotators; in the validated data the figures are 22.0% and 0.455 to 0.491, and review changed more than a third of the utterances. We define splits disjoint on the speaker identifier and tone-aware evaluation metrics, discuss what the measured tonal noise implies for training and evaluation, and outline the planned release of the data.

Keywords: Fongbe, speech corpus, low-resource speech recognition, tonal languages, tone annotation, annotation quality

1 Introduction

It is often said that speech recognition is a solved problem. It is solved for a small number of languages that have thousands of hours of transcribed speech. For the others it has barely started: public speech data exist for only 61 of Africa's more than 2,000 languages, and most of those 61 have less than one hour (Elmadany et al., 2025; Joshi et al., 2020).

Fongbe is one of the languages left out. Fongbe (Fon) is a Gbe language of the Volta-Niger family, spoken mainly in southern Benin and also in Togo and Nigeria, by about 8 millions speakers Laleye et al. (2016). It is tonal: pitch distinguishes words and grammatical constructions, and the national orthography marks tone with diacritics (Lefebvre and Brousseau, 2002). Many of its speakers read neither French nor written Fongbe. For them, speech is not one interface among several. It is the only one. A voice service that mishears them is a service they cannot use.

What is missing? Models are not what is missing. Self-supervised encoders pre-trained on hundreds of thousands of hours now cover more than a thousand languages (Babu et al., 2022; Pratap et al., 2024), and compact variants centred on African languages exist (Alabi et al., 2025; Boito et al., 2024). It is tempting to conclude that low-resource recognition has become a matter of fine-tuning a large model on a few hours of speech. The evidence says otherwise. Without data in the target language, the best multilingual recognizers average more than 75% word error rate on African test sets of the same survey. With twelve hours of read Fongbe, a fine-tuned model reaches 9.48% WER on its benchmark, and native speakers then rate its transcriptions of real-world audio at 36.5 out of 100 (Adjovi et al., 2026). The benchmark number is correct. It measures the wrong thing.

Our position is simple. The limiting factor for Fongbe speech technology is data, and more precisely three properties of data that no amount of pre-training on other languages will supply.

The first is scale in the language itself. Public Fongbe resources total a few tens of hours. The ALFFA corpus supported the first continuous recognizer with about ten hours of read speech (Laleye et al., 2016). FFSTC and FFSTC 2 reach 61 hours, paired with French translations (Kponou et al., 2024, 2025c). The large multilingual collections of recent years, such as WAXAL, do not include Fongbe (Diack et al., 2026). One cannot adapt an encoder to a language with data one does not have.

The second is speech as people actually produce it. Almost all available Fongbe speech is read aloud or produced from written prompts. Nobody asks a question about a land title, a weather forecast or a payment by reading a sentence from a screen. A system trained and tested on read speech learns read speech.

The third is tone, written reliably. A child acquires the tones of Fongbe from speech alone, years before seeing a diacritic. A recognizer has no such luck. It learns tone only from what annotators write, and annotators do not write it the same way: everyday written Fongbe often omits tone marks altogether. A model trained on such transcripts either drops tone, which leaves its output ambiguous, or learns an inconsistent target. The damage does not stop at training. A tone error rate measured against inconsistent references is not a measurement.

None of these three is an engineering detail. Each is a precondition for doing the science. One cannot compare tonal models without tone-marked speech, and one cannot trust the comparison without knowing how noisy the marks are. Section 2.2 places these obstacles in the wider literature.

This paper. We present a large-scale Fongbe speech dataset built to remove these three obstacles. It was collected within *Jaimemalangue*, a national program for Beninese languages, in which it supports both research and voice applications released as digital public goods. Our contributions are:

1. **A corpus of about 700 hours of Fongbe speech**, collected through a web platform from 2,420 registered contributors across Benin, with a stable identifier for every recording session. Speakers formulate their own speech, from photographs or written French sentences. They do not read Fongbe text.

2. **Tone-marked transcriptions with a documented review status**: about 351 hours transcribed, of which 159 validated hours, in the frozen export, were checked by a second person. The reliability of the rest is characterized rather than hidden.
3. **An audit of tonal annotation quality**, which quantifies tonal variation across word forms and across annotators, and its effect on tone-sensitive evaluation.
4. **A protocol for using the data**: splits disjoint on the speaker identifier and deduplicated, and metrics that score tone explicitly.

We do not claim that the corpus is clean. We claim something more useful: we know in what ways it is not, and we report them with numbers. Before review, a quarter of the written word forms occurred with more than one tonal spelling, and annotators differed by 31% in how densely they marked tone. After review, a fifth of the forms still do. A dataset paper that hides such facts produces benchmarks that reward the imitation of annotators. We would rather publish the noise and design the evaluation around it.

This paper describes data. It trains no systems, and quotes error rates only to audit the references: claims about systems belong with the benchmark, where they can be reproduced. It precedes the public release of the data. It gives the release, a planned benchmark and a planned shared task a citable description to rest on. Section 2 reviews related work. Section 3 describes how the data were collected and Section 4 the resulting corpus, its quality and the protocol for using it. Section 5 states the limitations, what they imply for users, and the ethical considerations.

2 Related Work

This section describes the speech resources that exist for African languages and for Fongbe (Section 2.1), states the five bottlenecks that limit them (Section 2.2), reviews the models built on them, with attention to tone and to unreliable labels (Section 2.3), and positions the present corpus (Section 2.4). Each corpus is cited once, in Table 1 or Table 2, and is referred to by name afterwards.

2.1 Existing resources

African languages. Africa has more than 2,000 languages, and Joshi et al. (2020) place most of them in the lowest classes of their resource taxonomy. Two recent surveys of African speech recognition name data scarcity as the first obstacle (Imam et al., 2025a,b). The corpora that exist fall into five families (Table 1).

Read and crowdsourced speech. NCHLT recorded about 200 speakers reading prompts in each of the eleven official languages of South Africa. Common Voice relies on volunteers who read and validate sentences. NaijaVoices used community-run recording of curated sentences for Hausa, Igbo and Yoruba. AfriSpeech-200 gathers African-accented English. FLEURS offers read speech in 102 languages, 20 of them Sub-Saharan, and is the usual multilingual test bed; it does not include Fongbe.

Religious and single-speaker recordings. BibleTTS provides studio recordings of one speaker per language reading the Bible. The labelled data behind the MMS models come from similar readings of the New Testament (Pratap et al., 2024).

Broadcast and radio archives. The West African Radio Corpus gathers untranscribed radio from Guinea. Zambezi Voice combines read speech with untranscribed radio for Zambian languages. Kunkado draws Bambara speech from Malian radio. Kallaama offers transcribed spontaneous speech on agriculture in Wolof, Pulaar and Sereer.

Prompted spontaneous speech. WAXAL elicited natural speech by asking speakers to describe images, and added read speech for synthesis. The present corpus belongs to this family.

Corpus	Languages	Volume	Speech type
ALFFA (Gauthier et al., 2016)	4, including Fongbe	about 59 h in total	Read, broadcast news
NCHLT (Barnard et al., 2014)	11 (South Africa)	over 50 h per language	Read
Common Voice, African subset (Ardila et al., 2020)	21	about 1,840 h, of which 1,354 h Kinyarwanda	Read, crowdsourced
FLEURS (Conneau et al., 2023)	102, of which 20 Sub-Saharan	about 12 h per language	Read
BibleTTS (Meyer et al., 2022)	10	up to 86 h per language	Read Bible, single speaker
West African Radio Corpus (Doumbouya et al., 2021)	10 (Guinea, incl. French)	about 144 h, untranscribed	Radio
Zambezi Voice (Sikasote et al., 2023)	5 (Zambia)	79 h transcribed, 525 h untranscribed	Read; radio
AfriSpeech-200 (Olatunji et al., 2023)	African-accented English	200 h, 2,463 speakers	Read
Kallaama (Gauthier et al., 2024)	3 (Senegal)	125 h	Spontaneous, agriculture
NaijaVoices (Emezue et al., 2025)	3 (Nigeria)	1,838 h, 5,455 speakers	Read
Kunkado (Diarra et al., 2025)	Bambara	161 h, of which 39 h corrected by humans	Radio
WAXAL (Diack et al., 2026)	24	about 1,250 h (ASR), 235 h (TTS)	Image-prompted spontaneous; read

Table 1: Representative speech corpora for African languages. Volumes for ALFFA and Common Voice are those counted by Elmadany et al. (2025).

Multilingual projects. ALFFA collected read and broadcast speech for Wolof, Fongbe, Swahili and Amharic.

Fongbe. Public Fongbe resources grew from about ten hours to a few tens of hours in ten years (Table 2).

The first continuous speech recognition system for Fongbe was built within the ALFFA project. Native speakers read about 1,500 everyday phrases, recorded at 16 kHz in a quiet environment with the LIG-Aikuma mobile application, for 10,402 segments and around 10 hours. Acoustic models were trained with Kaldi on MFCC features, from monophone GMM/HMM systems to subspace GMMs with discriminative training, and trigram language models with SRILM on about one million words of text from websites, Bible translations and other documents. The best system reached 44.04% WER on tone-marked text. OkwuGbé later trained end-to-end neural models for Fon and Igbo on this corpus (Dossou and Emezue, 2021). A second set of recordings with Fongbe transcripts was released on Zenodo in 2022.

FFSTC assembled about 31 hours of Fongbe speech paired with French text. It combines ALFFA recordings, sentences of the FFR translation corpus (Dossou and Emezue, 2020), and French sentences that participants translated aloud on a web platform. Its baseline end-to-end systems reached BLEU scores of 8.96 and 8.14. FFSTC 2 added 36 hours of read speech with Fongbe transcripts, for 61 hours in total, and compared cascade and end-to-end translation. The Fongbe-French pair entered the IWSLT 2025 low-resource track on this basis (Kponou

Resource	Fongbe speech	Speech type	Annotation
ALFFA Fongbe (Laleye et al., 2016)	about 10 h	Read	Fongbe transcription
Fongbe Speech Dataset (Laleye, 2022)	about 3,979 recordings	Read, 13 participants (public description)	Fongbe transcription
FFSTC (Kponou et al., 2024)	about 31 h	Read and translated prompts	French translation
FFSTC 2 (Kponou et al., 2025a,c)	61 h in total	Read prompts	Fongbe transcription and French translation
This work	about 700 h	Prompted by images and French sentences; formulated by the speaker	Tone-marked Fongbe transcription for about 351 h; 159 h validated by a second person

Table 2: Speech resources for Fongbe.

et al., 2025b).

2.2 Scientific bottlenecks

Five bottlenecks limit these resources. They are numbered B1 to B5 and reused in Section 2.4. For each one we give the evidence across languages first, then the situation of Fongbe. All figures are taken from the cited papers; shares and ratios are our own calculations from those figures.

B1. Too little transcribed speech, concentrated in a few languages. The transcribed speech available for African languages is small, unevenly spread, and below what current models are known to need.

(a) The total is small and concentrated. A recent survey counts 8,605 hours of public speech data for 61 African languages, of which 6,109 hours, in 42 languages, are transcribed for recognition (Elmadany et al., 2025). Five languages, Kinyarwanda, Hausa, Yoruba, Swahili and Igbo, hold 4,345 hours, half of the total, while 35 of the 61 have less than fifty hours and 16 less than one. Three corpora, NCHLT, NaijaVoices and Common Voice, supply 92% of the transcribed hours. For scale, Whisper was trained on 680,000 hours, of which 117,000 cover 96 languages other than English (Radford et al., 2023), and the labelled data of MMS reach 1,107 languages with an average of 32 training hours each.

(b) Volume does not follow the number of speakers. In the survey, Oromo has 34.6 hours for about 45 million speakers and Nigerian Pidgin 0.2 hours for about 120 million, while Swati has 307 hours for 4.7 million speakers: under one hour per million speakers in the first case, 65 in the last. The well-resourced languages owe their position to a single project. Nearly all the hours of Kinyarwanda come from Common Voice, and nearly all those of Igbo from NaijaVoices.

(c) The amount needed is known. Across the languages of Whisper, word error rate halves for every sixteen-fold increase in training data. Fine-tuning Whisper large-v3 on Kinyarwanda gives 12.51% WER with 50 hours, 9.82% with 200 hours and 7.14% with about 1,400 hours, and the authors place viable performance between 50 and 200 hours (Akera et al., 2025). A benchmark on 13 African languages shows the same shape: XLS-R on Swahili goes from 58.23% WER with one hour to 26.36% with 50 hours and 11.36% with 400 hours (Nahabwe et al., 2025).

(d) Fongbe. The survey counts 7.2 hours for Fongbe, which is the ALFFA corpus. With the 36 transcribed hours of FFSTC 2, the total is about 44 hours. Fongbe thus sits below the 50-

Language	System	Test set A	WER	Test set B	WER
Wolof	MMS	ALFFA, read	43.57	Kallaama, spontaneous	101.14
Wolof	Simba-M (fine-tuned)	ALFFA, read	34.75	Kallaama, spontaneous	100.44
Xhosa	MMS	NCHLT, read	35.24	Lwazi, telephone	73.89
Zulu	MMS	NCHLT, read	31.31	Lwazi, telephone	70.12
Afrikaans	MMS	Common Voice, read	26.29	Lwazi, telephone	92.06

Table 3: Word error rate (%) of the same system on two test sets of the same language. Values are those reported by Elmadany et al. (2025).

hour mark, on the steepest part of these curves. With this volume a language supports neither robust acoustic modeling nor a test set large enough to compare systems, and conclusions drawn on the few well-resourced African languages are extended to it without evidence.

B2. Narrow domains and narrow speaker bases. Most of the available hours are read aloud, from a narrow range of texts, by few or unbalanced speakers, and systems trained on them do not carry over to other conditions.

(a) Read speech dominates. The three corpora that supply 92% of the transcribed hours are read speech, and so is the main benchmark, FLEURS. The authors of MMS write that their Bible recordings are almost always single speaker, that speakers are often male, and that such data may lead to poor performance on other domains or on unseen speakers. The authors of AfriSpeech-200 warn that, since all their samples are read from prompts, models may underperform on spontaneous speech.

(b) Speaker bases are thin or skewed. An audit of Common Voice 17, FLEURS and VoxPopuli found that Zulu and Northern Sotho each come from one speaker in Common Voice (Lau et al., 2025). In Kallaama, women produce 17.3% of the speech overall and 10.2% of the Wolof. Bases of several thousand speakers, as in NaijaVoices and AfriSpeech-200, are recent exceptions.

(c) The effect is large, and it has been measured. Table 3 reports the same system on two test sets of the same language. MMS obtains 43.57% WER on read Wolof and 101.14% on the spontaneous Wolof of Kallaama, and fine-tuning does not close the gap. On Xhosa and Zulu, error doubles between broadband read speech and telephone speech. These gaps combine speech style, channel, segmentation and transcription conventions, and a benchmark built on one corpus controls none of them. The pattern is not specific to Africa: the Whisper study reports that models trained on LibriSpeech make about twice as many errors as humans on other English datasets. Text mismatch has the same effect: in the benchmark cited under B1, language-model decoding raises WER on Bemba from 49.63% to 60.72%.

(d) Fongbe. Every Fongbe corpus described in the literature is read or prompted from text, and speaker bases are small: 28 speakers in ALFFA, 8 of them women, and 20 participants aged 20 to 40 for the recordings added by FFSTC. FFSTC 2 does not report its number of speakers; each read at least 2,000 sentences translated from French Common Voice text, and speakers whose accent showed Mahi or Gungbe influence were excluded. One study indicates the consequence. A model that reaches 9.48% WER on the read ALFFA test set was applied to Fongbe videos found online, and native speakers rated its transcriptions at 36.5 out of 100 on average (Adjovi et al., 2026). That evaluation covers 50 segments and should be read as an indication, not as a measurement.

B3. Transcription is the limiting step. Collecting audio has become relatively easy. Turning it into reliable text has not.

(a) Audio outpaces transcription. WAXAL released transcripts for 10% of the audio its partners collected. Zambezi Voice has nearly seven untranscribed hours for each transcribed one, and the West African Radio Corpus has no transcripts at all. Kunkado started from about 300 hours of radio; a quarter of the 161 hours it released was corrected by humans, and the rest carries machine transcripts. In the survey, 1,960 hours in 32 languages are audio only.

(b) The cost per hour is high. In Kunkado, seven annotators spent about 1,260 hours correcting 39.3 hours of speech, 32 hours of work per hour of audio, although they started from an automatic transcript. In Kallaama, three students of linguistics needed nine months to transcribe the corpus. For newly documented languages, up to 50 hours of work per hour of audio are reported (Liang and Levow, 2025).

(c) The obstacle is the written language. The authors of Kallaama report that the official spelling rules are rarely used by the population, that keyboards lacked some of the required characters, and that transcribers sometimes could not understand the dialect of a recording, which forced a second collection campaign. The annotators of Kunkado were not held to a single Bambara orthography. Writers who follow a fixed convention are few, so annotators must be trained and checked, and code-switching with French, English or Arabic adds decisions that guidelines must settle. Writing tone is the hardest part (B4).

(d) The workarounds move the problem elsewhere. Prompted reading, where the text exists before the audio, avoids transcription at the price of the narrowness described under B2. Machine pre-annotation speeds the work: on Indian languages, post-editing took 1.2 minutes per utterance against 2.4 for transcription from scratch (Bhogale et al., 2026). It did not prevent the 32-to-1 ratio above. Automatic labels left unreviewed carry model errors into the data: in the Kikuyu part of the data-scaling study cited under B1, noisy or unclear references account for 38.6% of the utterances with high error.

(e) Fongbe. Among the corpora of Table 2, none contains spontaneous Fongbe speech transcribed by hand. In ALFFA and in FFSTC 2 the Fongbe text existed before the recording; in FFSTC 2 it was drafted by machine translation from French and revised by linguists. FFSTC pairs Fongbe speech with French text only. The spontaneous speech that matters most for applications is therefore the least transcribed, and participatory approaches that involve speaker communities have become the norm for building resources (Nekoto et al., 2020).

B4. Tone is under-specified in transcripts, models and metrics. Tone carries meaning in most African languages, and it is the layer that writing, corpora, models and metrics handle worst.

(a) Tone is the rule in African languages and optional in their writing. In the sample of the World Atlas of Language Structures, 220 of 527 languages are tonal and virtually all the languages of Africa are; of 68 Niger-Congo languages sampled, five are not (Maddieson, 2013). Orthographies differ in how much of it they write, from no marking to exhaustive marking, with reduced systems in between (Bird, 1999b). Practice departs further from the norm. In diacritized Yoruba text, 85% of words carry diacritics, and yet, with very few exceptions, diacritics are omitted from electronic texts (Orife, 2018).

(b) Writing tone is hard, even for trained writers. In an experiment on Dschang, a Grassfields language of Cameroon, nearly all participants had attended tone-marking classes. Tone marks degraded their reading fluency. Asked to add tone marks to unmarked text, experienced writers reached 83.5% accuracy and inexperienced writers 53%, which the author describes as not much better than chance (Bird, 1999a). The transcripts of a speech corpus are produced by writers of this kind.

(c) What is lost without tone can be measured. In the Yoruba study above, 32% of word types written without diacritics correspond to two or more diacritized forms, and 64% of

monosyllabic types do; a word without diacritics has 1.47 diacritized alternatives on average, against 1.04 for French. For Fongbe, removing tone marks from the ALFFA text reduces the vocabulary from 10,130 to 8,244 words, the out-of-vocabulary rate from 9.1% to 4.96% and the perplexity of the language model from 591 to 138. Tone-free text is easier to model because it merges distinct words. The loss shows downstream: in FFSTC 2, translating manual Fongbe transcripts into French gives 58.9 BLEU with diacritics and 47.39 without, and the authors conclude that diacritics matter more for translating Fongbe than for recognizing it.

(d) Corpora rarely say what they did. WAXAL states that transcriptions use the local script and gives no guideline on tone marks. NaijaVoices settled discrepancies in diacritics case by case. Kunkado removes diacritics before training. The builders of the Yoruba corpus Ìròyìn-Speech found diacritics missing in the news text they crawled and verified them word by word (Ógúnrèmí et al., 2023). FFSTC does not discuss tone marking. Transcripts may therefore carry full, partial or no tone marking, and the choice is seldom documented.

(e) Models and metrics hide the problem. Self-supervised encoders encode lexical tone, even when trained on non-tonal languages, and fine-tuning for recognition weakens that information in models pre-trained on non-tonal languages (Shen et al., 2024). Whether fine-tuning on targets that drop or scramble tone has the same effect has not, to our knowledge, been tested on an African language. On Yoruba and Uneme, tones remain the hardest features for current models, and a WER of 0.788 on Yoruba corresponds to a phonological feature error rate of 0.151 (Chen et al., 2026). Standard metrics do not expose this: WER counts a wrong tone as a wrong word, CER counts it as one character among many, and confidence scores do not flag it.

B5. Evaluation and documentation are weak. The results reported for African languages rest on test sets and reference transcripts that are rarely examined. We read the papers describing nine corpora, three of them for Fongbe, and recorded how each handles its test set, its partition and its references (Table 4). Several of the Fongbe resources were built by one or more authors of the present paper, and the criticism applies to them without exception.

(a) Test sets are small. FLEURS sets aside 350 sentences per language; its 20 Sub-Saharan languages share 58 hours of test audio, under three hours each. For Fongbe, the ALFFA test set comes from 4 speakers and the ASR test set of FFSTC 2 holds under four hours. AfriSpeech-200 shows that a different scale is feasible, with 750 test speakers.

(b) Partitions are not controlled, and results depend on them. Common Voice and AfriSpeech-200 guarantee that a speaker appears in a single split. This is not the rule. ALFFA Fongbe offers two partitions, and neither is disjoint on both axes: the category partition takes its test data from 4 speakers and its training data from 25, out of 28, and the speaker partition reuses nearly all phrases on both sides, with 1,421 phrases in training and 1,410 in test. In FFSTC, several participants could record under the same credentials. In FFSTC 2, duplicate transcripts were removed from validation and test only. The effect of such choices is measurable: on Fongbe, holding out one speaker at a time spreads word error rates over 54.74 points, with a standard deviation of 12.99, against 1.05 across random splits (Liu et al., 2023). This uncertainty is almost never reported.

(c) References are checked once, and agreement is not measured. WAXAL and Kallaama used a single transcriber per recording, and the other corpora rely on one round of votes or validation. None of the nine papers reports an agreement statistic between transcribers or validators. The cost of this omission has been measured elsewhere. In the Indian-language study cited under B3, two human references for the same audio differ by 20.4% WER on average. For English, scoring one system against several references instead of one lowers its WER from 11.21% to 5.45% (McNamara et al., 2024). A reported WER thus mixes recognition errors with disagreement among humans, in a proportion that the reader cannot estimate.

(d) The written norm is left implicit. The audit cited under B2 found that the Cantonese

Corpus	Test set	Speakers disjoint	Check of references	Agreement	Tone
ALFFA Fongbe	2,168 segments, 4 speakers	no (category partition)	read prompts	n.r.	n.r.
FFSTC	4 h, 2,482 sentences	n.s.	votes; 0–5 score	n.r.	n.r.
FFSTC 2, ASR	3.93 h, 2.5k sentences	n.s.	one pass by a pair of validators	n.r.	n.r.
FLEURS	350 sentences per language	yes (test)	crowd validation	n.r.	n.r.
Common Voice	varies by language	yes	two up-votes	n.r.	n.r.
AfriSpeech-200	18.77 h, 750 speakers	yes	up and down votes	n.r.	n.a.
NaijaVoices	3.34 h, 20 speakers	yes (test)	review of a random subset	n.r.	n.r.
WAXAL	not described	not described	one annotator, then quality control	n.r.	n.r.
Kallaama	none defined	–	one transcriber; partial expert review	n.r.	n.a.

Table 4: Evaluation and documentation practices as described in each corpus paper. Agreement: a statistic of agreement between transcribers or validators. Tone: a measure of the consistency of tone marking. n.r.: not reported; n.s.: not stated; n.a.: not applicable, tone is not written.

portion of FLEURS is 89.8% Standard Written Chinese and that its Egyptian Arabic portion is 98.6% Modern Standard Arabic. The Yoruba transcripts of FLEURS do not follow standard Yoruba orthography (Alabi et al., 2025). A benchmark then measures conformity to a norm that the corpus never states.

(e) Tone consistency is never quantified. This is the weakness that matters most for the present work. Removing tone marks lowered the WER of the 2016 Fongbe system by 29 points, to 14.83%. In FFSTC 2 the figures are 21.98% with tone marks and 17.02% without, a comparison its authors caution is not direct. For Yoruba, Whisper-small reaches 65.70% WER on tone-marked transcripts and 37.90% on the same transcripts without tone marks (Olusanya, 2026). None of these studies measures how consistently the reference tone marks were written. Where agreement on tonal labels has been measured, it is imperfect even with trained transcribers and a mature convention: 81% pairwise agreement on the exact label for English intonation under ToBI (Pitrelli et al., 1994). The gap between tonal and tone-free WER therefore confounds the errors of the model with the inconsistency of the references.

(f) Access and basic figures. The FFSTC papers state no licence, and the IWSLT 2025 release was restricted to the shared task, with test data in a private repository (IWSLT, 2025). Basic figures also differ between sources: FFSTC 2 is 61 hours in its paper and nearly 57 on the shared task page, and ALFFA Fongbe is around 10 hours in its paper against the 7.2 counted by the survey.

Sections 4.5 and 4.7 respond to these points for the present corpus and state where the response is still incomplete.

2.3 Models, tone and unreliable labels

Multilingual pre-training. Multilingual pre-training reduces the amount of transcribed speech a language needs. XLS-R was pre-trained on 436,000 hours in 128 languages (Babu et al., 2022), and MMS on 491,000 hours in more than 1,400 languages. mHuBERT-147 showed that a compact HuBERT model trained on 90,000 hours in 147 languages, 16 of them African, can match much larger encoders (Boito et al., 2024; Hsu et al., 2021). Two efforts target Africa: a

HuBERT model pre-trained on about 60,000 hours from 21 Sub-Saharan languages (Caubrière and Gauthier, 2024), and AfriHuBERT, which continued the pre-training of mHuBERT-147 on more than 10,000 hours covering 1,226 African languages, many of them with less than ten hours. Supervised multilingual recognizers such as Whisper, SeamlessM4T (Seamless Communication, 2023) and Omnilingual ASR (Omnilingual ASR Team, 2025) extend coverage further.

Pre-training does not remove the need for data in the language. In the survey discussed under B1, MMS, SeamlessM4T and Whisper average more than 75% WER across 56 African test sets, and fine-tuning on five hours per language brings large gains. For Fongbe, only MMS lists the language as supported. On the ALFFA test set, the three models exceed 85% WER without Fongbe fine-tuning, and the best multilingually fine-tuned model reaches 33%. Coverage in pre-training prevents complete failure; accuracy is decided by in-language data and by its match to the target domain. Continued pre-training on untranscribed speech of the target language is the natural use of audio that cannot be transcribed (B3), provided reliable transcripts exist to fine-tune and to evaluate.

Modeling tone. Research on tone in speech recognition has developed mostly for Mandarin, Cantonese and Vietnamese, where large tone-labelled corpora exist. Classical systems added pitch features to the acoustic front end and used tone-dependent phones (Ghahremani et al., 2014). End-to-end work treats tone as a sequence prediction problem that a network learns from the signal, for example with CTC (Lugosch and Tomar, 2018). Recent work acts at three levels. In the encoder, SITA adapts wav2vec-style models so that representations become invariant to the speaker while keeping lexical tones apart, on Hmong and Mandarin (Xu et al., 2026). In the output layer, Latent Softmax models tone-marked vowels as subclasses of their base vowels inside a CTC-compatible layer, so that tonal and non-tonal supervision can be mixed (Jiang et al., 2026). After recognition, CantoASR pairs a Whisper model adapted for tone discrimination with an audio language model that uses prosodic cues to correct the transcript (Chen et al., 2025). For African languages the dominant approach is restoration from text: the Yoruba study cited under B4 restores diacritics with 95.4% accuracy using a sequence-to-sequence model, and Yoruba speech corpora have been built with diacritized transcripts (Gutkin et al., 2020).

Learning from unreliable labels. A corpus with a reliable tier and a noisier one calls for methods that tolerate label errors. Semi-supervised training with pseudo-labels, as in noisy student training, uses a model trained on trusted data to label the rest (Park et al., 2020). Other methods change the training criterion. Star Temporal Classification extends CTC to transcripts with missing tokens and recovers most of the supervised performance when up to 70% of the labels are missing (Pratap et al., 2022). Omni-temporal Classification handles substitutions and insertions as well and remains stable with transcripts containing up to 70% errors (Gao et al., 2023). These results were obtained on English with simulated noise. Whether they hold for real annotation noise concentrated on tone marks is an open question that requires a corpus where such noise is present and measured.

Open questions. Three questions follow, and none can be answered without a large tone-marked corpus. Should tone be predicted jointly with segments, by a separate head, or restored afterwards from text? How much of the tone in a transcript is predictable from text alone, and how much needs the audio? And how should systems be compared when the reference tone marks are themselves inconsistent (B5)?

2.4 Positioning

Table 5 states how the present corpus addresses each bottleneck, and where it does not.

Bottleneck	Contribution of this work	What remains open
B1. Scarce transcribed speech	About 700 h of Fongbe speech, about 350 h transcribed; an order of magnitude above prior public resources	One language; hours per tier still provisional
B2. Narrow domains and speakers	Guided spontaneous speech from more than 2500 speakers, with speaker identifiers	No telephone or conversational speech; coverage by region to be reported
B3. Transcription as limiting step	Trained annotators on a dedicated platform; a lower-reliability tier kept and labelled instead of discarded	About half of the audio is untranscribed
B4. Tone under-specified	Tone-marked transcripts, a written convention, tone-aware metrics, three tonal modeling baselines	Convention still being harmonised
B5. Weak evaluation and documentation	Audit of tonal consistency; speaker-disjoint, deduplicated splits; harmonised test set; datasheet	Inter-annotator agreement not yet measured

Table 5: Bottlenecks identified in Section 2.2 and how this work addresses them.

3 Data Collection Methodology

A corpus is a set of decisions made before anyone presses record. We made four, and everything below follows from them.

3.1 Design goals

- **Scale in the language itself.** Enough untranscribed speech to adapt a self-supervised encoder to Fongbe acoustics. Pre-training on other languages does not replace it.
- **Tone in the transcripts.** Transcriptions that write tone. A model cannot learn what the labels do not contain.
- **Speaker identity.** A stable identifier for every recording session, so that splits can be disjoint on it. Without it, test scores measure memorised voices.
- **Speech people produce.** Speech that speakers formulate themselves, not text they read aloud. Users do not read sentences to a voice service. They talk to it.

None of the four is new. Each answers a failure of earlier practice that Section 2.2 documents: thin data (B1), read speech (B2), single-pass transcription (B3), tone left unspecified (B4), and partitions that leak speakers (B5). The subsections below say what earlier corpora did, where that breaks, and why we did otherwise.

3.2 Project, roles and workflow

The data come from the *J'aime Ma Langue* project, which collects voice data in Fon (Fongbe). It is run by the Institute for Inclusive Digital Africa (IIDiA) with the Beninese Agency for Information Systems and Digital Affairs (ASIN). We organised the work as a pipeline of five roles (Table 6), each with a written working guide. The principle is simple: no item enters the validated data until a person has listened to the audio and a second person has checked the text against it. A linguistic supervisor (*encadreur linguistique*) settles disputes and answers questions on the language.

Role	People	Task	Decision
Contributor (collector)	2,420	Records a spoken response to a prompt	Submits the recording
Audio validator	45	Listens to each recording from start to end	Accepts or rejects the audio; logs anomalies in a register
Transcriber	31	Writes what is said, with tone marks	Submits the transcription
Transcription validator	11	Checks the text against the audio, line by line	Corrects minor errors and validates, or rejects
Linguistic supervisor	1	Answers language questions and reviews disputed items	Final accept or reject

Table 6: Roles in the collection and annotation pipeline, as defined in the working guides, with the number of people who held each role.

Annotators (validators and transcribers) were recruited through an open call on social networks, and trained in five one-hour sessions on the rules for writing Fongbe. They were paid according to objectives, in hours of audio processed.

A caveat that matters. The guides say what people were told to do. They do not say what people did. How closely the transcripts follow the rules is an empirical question, and Section 4.5 answers it.

3.3 Collection platform and participants

All recordings of the project were collected through a web platform operated by the project team, where contributors register, possibly with a telephone number, and record on their own phone, tablet or computer. Audio is captured at 48 kHz as mono Opus in a WebM container, and is converted to 16 kHz mono FLAC for release.

Contributors are people who can express themselves in Fon; that was the selection criterion, and it is declared, not tested. In all, more than 2,500 contributors registered. They come from the twelve departments of Benin (Atlantique, Littoral, Collines, Borgou, Ouémé, Alibori, Zou, Atacora, Donga, Mono, Plateau and Couffo), with registrations from towns such as Abomey, Ouidah, Cotonou, Parakou, Kandi, Malanville, Djougou, Kouandé and Glazoué. We do not claim a dialect-balanced sample: no dialect label was collected. Each speaker gave informed consent before recording (see Section 5.3).

Roles were kept apart, with small exceptions that we report. Over the 318,432 recordings, one contributor transcribed their own recording, none validated their own audio or transcription, and ten transcriptions were validated by the person who wrote them. Thirteen people held both the contributor and transcriber roles, and five both the contributor and transcription-validator roles, on different recordings.

3.4 Elicitation

Contributors answered prompts shown on screen, and each prompt is an image. There are 3,823 distinct images of three kinds (Table 7). About half of the recordings answer a photograph of everyday life or of a theme (hotels, market, banks, children, supermarket, house, cars, money, mobile money), which the contributor describes in their own words. The other half answer a written French sentence, shown as an image, which the contributor says in Fongbe. The platform asks for a response of 5 to 15 seconds.

The audio validators' guide defines what a usable response looks like. For a description, a contributor must:

Prompt	Distinct images	Recordings	Share
Written French sentence, to say in Fongbe	2,126	159,312	50.0%
Photograph of everyday life	1,123	157,601	49.5%
Thematic photograph	574	1,519	0.5%
Total	3,823	318,432	100%

Table 7: Prompts used to elicit speech on the platform.

- speak naturally and continuously, and not read a text aloud;
- avoid fillers such as *euh* and *hmm*;
- avoid repeating a word twice in a row, which mostly means stuttering, and avoid repeating the same idea;
- avoid phrases that point at the picture, such as "this image shows" or "I see", and describe the content directly;
- be as descriptive or creative as possible.

Why images, and not Fongbe text. The usual way to get a transcript is to write the text first and have people read it. That is how NCHLT, Common Voice, ALFFA and FFSTC 2 were built (Section 2.1). For Fongbe it has a specific flaw. The text must exist first. In FFSTC 2 it was drafted by machine translation from French, and the speakers must then read a language many of them rarely read. Reading also gives read prosody, and read speech is not what a voice service hears (B2). An image asks for no written Fongbe at all. The idea has a long history in speech and vision research, where spoken captions of Flickr images were collected by asking crowd workers to describe them (Harwath and Glass, 2015), and WAXAL applies it to African languages (Diack et al., 2026). What we add is the pairing of such prompts with human transcription written with tone marks in Fongbe, and a second, cheaper prompt, a French sentence to say in Fongbe. That second prompt fixes the content, but not the Fongbe words, so speakers still choose their own phrasing. It may also bring French constructions into the Fongbe, a possibility we have not tested. To our knowledge no earlier Fongbe corpus combines these ingredients; the public Zenodo set is read speech by 13 participants.

The result is natural phrasing and a broad vocabulary. It also has two costs, and we state them plainly. First, many speakers answer the same prompts, so utterances can share a structure and differ in a few words. This inflates test scores unless the splits account for it (Section 4.7). Overlap between training and test data makes a model look as if it generalises when it memorises, and it has been found in several public NLP benchmarks (Elangovan et al., 2021). Speech corpora built from a fixed set of prompts are exposed to the same risk. Second, disfluent speech was rejected by design. The corpus is therefore closer to fluent, planned speech than to conversation (Section 5). Models do worse on disfluent speech, and stuttered speech shows it most clearly (Mujtaba et al., 2024). Even on English conversational speech, a recognizer at human-level error still confuses a filled pause with a backchannel, a mistake humans rarely make (Stolcke and Droppo, 2017). Our corpus will not teach that. We chose this trade-off knowingly. It buys clean supervision and gives up some realism.

3.5 Audio validation

Why named validators, and not votes. The reference design for crowdsourced speech is Common Voice, in which anonymous volunteers vote on each clip (Ardila et al., 2020). It scales to thousands of hours. Its validation rests on votes, and a vote carries a verdict, not a reason,

so one cannot easily ask why a clip was rejected or whether validators drifted. We chose a smaller design: named validators with a written guide, a register in which every anomaly is logged with its cause, and a supervisor who settles disputes. It scales less, and it yields data about the filtering itself.

The platform first rejects some recordings automatically, for a level that is too loud or too quiet, a silence of three seconds or more, or a processing error. Validators then listened to the remaining recordings in full, preferably with headphones. They did not edit or trim the audio. Their job was to judge it, not to repair it. A recording is accepted when it is clear, mostly in Fon and usable without correction; the guide's example is a clear recording that is 95% Fon. It is rejected when one of the following holds.

- **Language**: the dominant language is not Fon. Code-switching, loanwords and other languages (French, Goun, Yoruba and others) are tolerated when coherent with the context, and are noted.
- **Signal**: cuts, echo, unacceptable volume, strong background noise, background music, or several indistinct voices. Minor irregularities such as breath or distant noise are logged, and the recording is kept if it stays usable.
- **Delivery**: recited, repeated or too hesitant speech; a volume that is not constant; or pronunciation that is not representative of current spoken Fon.
- **Content**: a response unrelated to the image, a description that breaks a rule of Section 3.4, or an inaudible segment.

Validators also logged atypical items, such as a loanword, an unusual tone or a rare expression, and reported recurring problems to the supervisor. The outcome is reported in Section 4.4.

3.6 Transcription and tonal annotation

Why two passes. Among the corpora we reviewed in Table 4, the common practice is one pass: a single transcriber, or one round of votes. Nobody then knows how often the first person is wrong. We put two people on every validated item: a transcriber, then a validator who checks the text against the audio. The platform's history confirms that validators edit the transcriber's text: each of the 91,155 modifications keeps the previous text and the new one. The validator reads the transcriber's text, so this is a check, not an independent transcription, and it does not yield an agreement statistic (Section 4.5).

Transcribers listened to each recording and typed what they heard in the national orthography, with tone marks. They were advised to use a keyboard with the Fongbe alphabet and diacritics. The rules were short: write what is said, do not interpret or rephrase, check before submitting, and skip an item when in doubt. The transcribers' guide also asks for a French translation of each sentence, but the exported data contain none (Section 5).

Transcription validators then read each line against the audio. They corrected minor errors in place (diacritics, typography, spelling consistency) and rejected a transcription with larger errors. The guide offers a rule of thumb: more than three errors, reject; fewer, correct and validate. A speaker who cannot be understood is flagged with the zone and the reason, and the item is invalidated. Recurring decisions go into a file shared by all validators and the supervisor. Review is not complete: on 13 October 2026, 53.4% of the transcriptions produced had been reviewed, and only reviewed ones are validated (Section 4.4).

Validated and unreviewed transcriptions. We keep two states apart throughout. A transcription is *validated* when a transcriber wrote it and a second person checked it. It is *unreviewed* when no second person has checked it yet, and its reliability is lower or unverified.

Mark	Vowels	Share of marks
Acute	568,986	78.7%
Caron	139,239	19.3%
Grave	14,556	2.0%
Circumflex	189	<0.1%
Macron	3	<0.1%
All marks	722,973	100%
No mark	819,885	—
All vowels	1,542,858	—

Table 8: Tone marks on vowels in the validated data (1,085,290 words).

Between the manifests of July 2026 and the frozen export of 1 October 2026, transcription validators corrected the text of many utterances in the platform. We call this the correction pass, and Section 4.5 reports its effect.

3.7 Orthographic and tonal conventions

Transcriptions follow the Beninese national alphabet: the Latin script extended with ɖ, ɛ and ɔ, and digraphs for single consonants, including *gb*, *kp*, *ny*, *hw* and *xw*. Nasal vowels are written with a following *n*.

Fongbe has two underlying tones, high and low, with rising, falling and mid realizations on the surface (Lefebvre and Brousseau, 2002). Tone is written with diacritics on the vowel. Table 8 counts the marks in the validated data. Acute and caron account for 98% of them. The grave is rare, which suggests that low tone is mostly left unmarked, and more than half of all vowels carry no mark.

The transcribers' guide asks for tone marks because they carry meaning: leaving them out can change the meaning of a word. But it stresses high and modulated tones, and says nothing about how completely every vowel must be marked, nor what an unmarked vowel means. That gap is a plausible source of the differences in tone-mark density between annotators that we measure in Section 4.5. The audit alone cannot prove it. No single written convention existed during annotation, and the guides are the only written rules. A convention note, validated by linguists, is planned and will be released with the data.

Word boundaries are a second source of variation, and the audit found examples such as *nyɔnu* 'woman' against *nyɔ nu* 'to be good for', and *mimɔ* 'the act of seeing' against *mi mɔ* 'you see'.

The guides add four rules that bear on recognition. The fourth, no interpretation, answers a specific failure. The audit cited under B2 found that the Cantonese portion of FLEURS is almost entirely Standard Written Chinese, and its Egyptian Arabic portion Modern Standard Arabic. The transcripts followed the written norm, not the speech. A benchmark built that way scores conformity to the norm.

- **Full word forms.** Unusual or excessive contractions are not written. The guide's example is *xwe dó é* instead of a contracted *xwe d'é*. Whole forms are easier for a recognizer to learn. Contractions that are common in speech remain allowed.
- **One spelling per word.** Where variants exist, the project picks a standard form and documents the choice.
- **Foreign words kept.** Code-switched words and loanwords, French in particular, are written as said and never translated.

- **No interpretation.** The transcript records what is said, not what the speaker meant.

3.8 Text normalization

Two layers of text exist, and users should not confuse them.

Published transcripts. They keep the text as the annotators wrote it. French words are in straight double quotation marks, which occur in 16.5% of utterances. There is no markup and no separate field. Proper names start with a capital letter, and about 95% of utterances begin with a capital and end with a full stop. The norm asks for numbers in letters, as spoken; digits still occur in 0.6% of utterances (336 of 60,953).

Training and scoring text. The pipeline used so far lowercases the text, removes punctuation and digits, and keeps tone marks. A tone-free projection is obtained by removing the marks. The scoring script that we will release adds four rules, so that scores can be tied to one exact text form.

1. **Unicode form.** Text is stored in NFD, so that a base letter and its tone mark are separate code points. NFC is for display only.
2. **Atomic digraphs.** *gb*, *kp*, *ny*, *hw* and *xw* are single units in tokenization and in character-level scoring.
3. **Closed inventory.** A linter rejects any character outside the Fongbe inventory and a short list of allowed symbols. Rejected items are flagged, never corrected silently.
4. **Projections.** Tone-free scores always come from the projection of the canonical form, never from a separately edited text.

3.9 Storage and versioning

Audio and transcripts carry stable utterance and speaker identifiers and are exported to a standard dataset format with one manifest per split. The data used in this paper are the split v2.1, built on 1 October 2026 from the export of the same date, and every manifest has a SHA-256 fingerprint. The split can be rebuilt byte for byte. The normalization script has no version number yet, and we count that among the limits of this release.

4 Dataset Statistics

4.1 Size and review status

The project collected about 700 hours of Fongbe speech. On 13 October 2026, about 351 hours of it had a transcription, and about 344 hours had none. Of the transcribed hours, 176 were validated by a second person. Table 9 gives the subsets used in the rest of the paper.

Table 10 gives the exact counts of the frozen split v2.1. Of the 159 validated hours, 131 were kept in the split. The rest was removed because its text duplicated development or test text (1,189 utterances) or because the speaker fell outside the partition.

4.2 Speakers and metadata

Two numbers describe the speakers, and they differ. In all, 2,420 contributors registered. The manifests of the validated data carry 710 speaker identifiers. An identifier corresponds to one recording session, so a person who recorded in several sessions holds several identifiers.

Subset	Hours	Transcription	Main use
All speech collected	700	Partly transcribed	Self-supervised adaptation
Transcribed (13 Oct 2026)	351	Tone-marked; 53% reviewed	—
validated by a second person	176	Tone-marked, checked	Supervised training, development, test
not yet validated	175	Unreviewed (about 166 h) or rejected (about 9 h)	Noisy supervision, robustness studies
Validated, frozen export (1 Oct 2026)	159	Utterances of 2 to 30 s	Source of the split
of which in split v2.1	131	After removing text duplicates	Training, development, test
Untranscribed	about 344	None	Self-supervised adaptation, pseudo-labelling

Table 9: Subsets of the corpus. Hours are rounded. The totals come from different dates (700 h from the audio audit of 1 October, the others from the platform registers of 13 October, except the export), so the untranscribed figure is an estimate.

Subset	Utterances	Hours	Speaker IDs	Words
Train	50,323	107.28	612	895,390
Development	4,248	9.58	38	72,252
Test	6,382	13.97	60	117,648
of which test-strict	5,020	11.01	44	—
Total, split v2.1	60,953	130.82	710	1,085,290

Table 10: Counts per subset. Words are occurrences after removing punctuation and digits, tone marks kept. A speaker ID is one recording session (Section 4.2).

Disjointness between splits is therefore guaranteed at the level of the identifier, not of the person.

Contributions are very uneven. The median identifier holds 1.2 minutes of audio, the smallest 0.08 minutes and the largest 3.75 hours. In training the median is 0.8 minutes, while development and test identifiers hold more, at 12.5 and 9.9 minutes. Sex and age are not in the manifests of the validated data. The metadata schema is given in Appendix **??**.

4.3 Lexical and tonal content

The content is what the protocol made it (Section 3.4). Prompts are images and written French sentences, so the vocabulary is concrete, and many speakers answer the same prompts. Fillers, stuttering and picture-pointing phrases were filtered out at validation. Loanwords and code-switching, mostly French, were kept when coherent with the context: French words appear in 16.5% of utterances (16.6% in training, 17.2% in development, 15.9% in test). The transcripts therefore contain some non-Fongbe tokens, and that is deliberate: it is how people speak.

Utterances are short (Table 11), as the platform asks for 5 to 15 seconds and the export keeps 2 to 30. The median is 7.1 seconds in the whole set and 7.4 in the test set.

The vocabulary has 13,778 distinct word forms with tone marks and 10,223 without, which is 1.35 times more forms when tone is written. In training alone the figures are 12,325 and 9,187. Table 8 gives the distribution of tone marks.

Repetition is limited but not absent. In the validated data, 549 texts occur more than

Subset	Min	Q1	Median	Q3	Max
Train	2.67	6.06	7.06	8.64	27.01
Development	5.02	6.24	7.44	9.42	21.82
Test	4.31	6.15	7.38	9.10	16.60
All	2.67	6.09	7.10	8.75	27.01

Table 11: Utterance duration in seconds (mean: 7.73).

once, which is 1,245 utterances (2.04%), and 216 of these texts are said by different speakers. Within the same speaker, 377 groups of identical utterances leave 439 utterances in excess (0.80 hours).

4.4 Validation outcomes

Audio and transcription validators logged accept and reject decisions with reasons (Sections 3.5 and 3.6). The registers, as of 13 October 2026, give the following.

Audio. Of 316,793 recordings processed (693.80 hours), 272,370 were accepted (590.21 hours), which is 86.0% of recordings and 85.1% of hours, and 44,423 were rejected (103.60 hours), which is 14.0%. Two thirds of the rejections are automatic: 30,213 recordings (71.79 hours, 68% of rejections) for a level that is too loud or too quiet, a silence of three seconds or more, or a processing error. Validators rejected 14,210 recordings (31.81 hours). They ticked one or several reasons: background noise or interference, poor audio quality, a problem of language or pronunciation, content unrelated to the image, a technical problem such as distortion or clipping, a level too low or too high, or a free-text reason.

Transcription. Annotators produced 163,974 transcriptions (351.11 hours). Of these, 87,629 (53.4%) had been reviewed: 84,062 were validated (176.44 hours) and 3,567 were rejected (8.72 hours), which is 4.1% of the reviewed transcriptions. The remaining 46.6% are not yet reviewed and are not validated.

4.5 Annotation quality audit

Tone is the least consistent dimension of the transcriptions. We measured by how much, twice: on the first split we built (v1.0), before review, and on the validated data (v2.1). The audit ran automatic checks over the transcribed data: Unicode and inventory linting, tonal variants per word form, tone-mark density per annotator, speech-rate outliers, language identification of the text, and near-duplicate detection. Agreement between three recognition systems was also used to flag suspect transcriptions. The three systems share training data, so their votes are not independent. We treat them as a filter, not as ground truth. The same caution applies to self-training with a teacher model, whose errors reappear in what it labels (Section 2.3).

Four observations follow from Table 12.

Tonal variation is concentrated. A fifth of the word forms appear with several tonal spellings, and these forms carry almost every occurrence. They are the frequent words. Reviewing the most frequent forms first therefore removes most of the variation at limited cost.

Annotators followed different conventions, and review narrowed the gap. Before review, tone-mark density differed by 31% in relative terms between the sparsest and the densest annotator, and four annotators marked about 0.38 to 0.39 marks per vowel. After validation

Finding	Before review (v1.0)	Validated (v2.1)
Tone-free word forms written with more than one tonal spelling	25.6% (4,330 forms)	22.0% (2,245 of 10,223)
Share of word occurrences carried by these forms	96.8%	94.5% (1,025,064 of 1,085,290)
Tone-mark density across annotators (marks per vowel)	0.377 to 0.495 (18 annotators)	0.455 to 0.491 (22 annotators)
Gap between WER with and without tone marks, same system (MMS), in points	11.5 (unreviewed references)	6.1
Share of that gap due to words that differ only in tone, pilot system	about 89%	not measured

Table 12: Main findings of the annotation quality audit. Density is computed for annotators with at least 2,000 vowels. The two manifests do not share annotator identifiers, so the comparison is between distributions, not between persons.

the difference is 8% and no annotator is below 0.40. The variation was systematic, not random. A model can then lower its tone error by imitating the majority annotator instead of listening to the speech. The same thing has been shown in text. In three language-understanding datasets, a few prolific crowd workers wrote most of the examples; giving the model the annotator's identity improved its score, and models often failed on examples from annotators they had not seen (Geva et al., 2019). Nothing suggests speech transcription is immune. The review is consistent with validators aligning the text, but the two manifests differ in content, so we do not claim it as the cause.

Tone dominates the tone-sensitive error. On a pilot system, nearly nine tenths of the difference between WER with and without tone marks came from words that are right except for tone. Part of that difference comes from the references, not the model. On the validated references the same gap is nearly halved, from 11.5 to 6.1 points, for one system. Validation corrects a large part of the tonal inconsistency, and does not remove all of it.

Review changes a lot of text. Between the manifests of July 2026 and the export of 1 October, transcription validators made 14,558, 21,754 and 25,859 modifications in July, August and September. Among the utterances that were already in the July manifests, and comparing with their July text without punctuation or case, 35.2% changed in training (10,820 of 30,750), 41.0% in development (1,136 of 2,772) and 38.8% in test (1,293 of 3,330). In training, the changes are in letters for 6,549 utterances, in tone alone for 2,453 and in word segmentation alone for 1,818, one category per utterance. Tone alone is therefore a fifth of the changes and 8% of the training utterances. More than a third of the unreviewed text was wrong in some respect, and any score computed on July references is not comparable with one computed on the current ones.

Not all tonal variation is noise. Before calling all of this noise, recall that human label variation is often treated as noise when part of it is signal (Plank, 2022). Grammatical tone, tonal sandhi and context legitimately change the tone of a word form. Telling a legitimate variant from an inconsistency is a linguistic decision, not a statistical one, and the project's linguists are reviewing it. Treating each annotator's labels as their own task, instead of collapsing them by majority vote, can match or beat training on the majority label on subjective tasks (Davani et al., 2022). That is why the release keeps annotator identifiers.

The audit also looked for other classes of problem in the validated data. Utterances whose speech rate is implausible for their transcript (fewer than 2.6 or more than 12 letters per second) number 387 in training (0.97 hours), 54 in development (0.13 hours) and 71 in test (0.15 hours). We did not measure transcripts that are not in Fongbe. Near-duplicate sentences across speakers and repeated sentences within a speaker are counted in Section 4.4, and the duplicates of development and test were removed from training when the split was built. No single convention note existed, and the review narrowed the gap between annotators without bringing them to one convention.

We have not yet measured inter-annotator agreement on doubly annotated data. This is the main gap, and it is a real one. There is no blind double annotation. The only measure available is the review of July to September, which is not an independent annotation. Professional human transcribers of English conversational speech disagree with the reference at 5.9% on Switchboard and 11.3% on CallHome, and a more careful protocol brings those to 5.1% and 6.8% (Stolcke and Droppo, 2017). These are words in a language with a settled orthography. With optional tone marks we have no reason to expect better. A double annotation of a sample of the test set is planned and will be reported in the next version of this paper.

4.6 Consequences for the release

The audit changes how the data should be used, and the release reflects it:

- **Training data are released as they are**, with review-status labels and annotator identifiers, so that users can model or filter the noise.
- **The test set needs more work.** Its utterances are validated by a second person, like the training ones, but have not been harmonised under a single written convention. We plan to have a small group of trained annotators re-listen to it, harmonise it and clean it of the classes above. The annotators of the current test set are not disjoint from those of the training set: 25 of its 28 annotators also transcribed training utterances, the other three transcribed one test utterance each, and all 24 annotators of the development set also transcribed training utterances. Keeping test annotators apart from training annotators is what the text-based evidence recommends (Geva et al., 2019); the planned group will be chosen for that. Speakers, as identified in the manifests, are disjoint across the three sets.
- **The validated data are described as validated, not as high quality.** Four measured facts support its reliability: a second person validated every transcription; the gap in tone-mark density between annotators narrowed; no development or test text appears in training; and the tone-sensitive gap in WER is nearly halved on validated references. Three facts limit the claim: there is no written tonal convention, no inter-annotator agreement is measured, and annotators are shared across the sets.
- **A list of accepted tonal variants** may accompany the evaluation script if linguists confirm that some forms legitimately admit several tones.

4.7 Splits and scoring conventions

We propose the following protocol for using the data. It follows from the audit, not from habit.

Splits. Training, development and test sets are disjoint on the speaker identifier: no identifier appears in more than one set. As Section 4.2 explains, an identifier is a recording session, so disjointness holds at that level. The test set holds 13.97 hours, 6,382 utterances and 60 identifiers; a stricter subset holds 11.01 hours, 5,020 utterances and 44 identifiers. Many

speakers answer the same prompts, so a random split would let models score well by memorising prompts and phrasings. The split removes from the training data any sentence identical, or within one word, to a development or test sentence (1,189 utterances), and keeps a single occurrence of each sentence per speaker in the test set. Development and test sets come from the validated data only.

Metrics. All metrics are computed on normalized text (Section 3.8) and micro-averaged over the test set, with macro-averages over speakers alongside to expose speaker imbalance. We define five.

- **CER**: character error rate, with digraphs and tone marks counted as units.
- **WER**: word error rate with tone marks. A word with one wrong tone counts as wrong.
- **WER-nt**: word error rate on the tone-free projection of reference and hypothesis.
- **TSER**: tonal syllable error rate. Reference and hypothesis are segmented into syllables, and a syllable is correct only if both its segments and its tone are correct.
- **Conditional TER**: among aligned syllables whose segments are correct, the proportion whose tone is wrong.

With S, D and I the numbers of substituted, deleted and inserted syllables and N_{syl} the number of reference syllables:

$$\text{TSER} = \frac{S + D + I}{N_{\text{syl}}}, \tag{1}$$

$$\text{TER}_{\text{cond}} = \frac{\#\{\text{syllables with correct segments and wrong tone}\}}{\#\{\text{syllables with correct segments}\}}. \tag{2}$$

WER and CER weigh tone poorly. In WER one tone error invalidates the whole word, and in CER it weighs a single mark whose cost depends on the Unicode form. TSER and conditional TER make tone explicit, and each guards against gaming the other: omitting hard passages to lower TER is penalised by TSER. The syllabification rules are being validated on the corpus and will be released with the scoring script.

5 Limitations and Considerations

A dataset paper should say where its data fail before others find out. This section lists the limits, then what they imply for anyone who trains or evaluates on the corpus, then the ethical questions.

5.1 Limitations

The main limitation is that we have not measured inter-annotator agreement on tone on doubly annotated data. The others follow.

- **No recognition results.** This paper describes data. It reports no baselines, and quotes error rates only to audit the references. Baselines and a benchmark will accompany the release, where the scoring script and the frozen manifests can be cited together.
- **Provisional review status.** Only 53% of the transcriptions produced are reviewed, so the validated hours will grow, and the figures of this paper will change with them.
- **Approximate figures.** Hours are rounded until the manifests are frozen for release.

- **Coverage.** The corpus consists of prompted responses. It contains no telephone or conversational speech. No dialect label was collected, and sex and age are absent from the manifests, so we cannot report its balance across dialects, age groups and sexes.
- **Fluent, guided speech.** Validators rejected hesitant, recited, repeated and filler-laden recordings by design, so the corpus under-represents disfluent and conversational speech. Models trained on it may degrade on such speech.
- **No translations.** The transcribers' guide asks for a French translation of each sentence, but the exported data contain none, so the corpus does not support speech translation as it stands. About half of the recordings answer a written French sentence, which gives a French source text for them.
- **Speaker identifiers are sessions.** A person who recorded in several sessions holds several identifiers, so splits are disjoint on the identifier, not necessarily on the person. Contributions are also very uneven: the median identifier holds 1.2 minutes of audio.
- **Shared annotators.** The annotators of the test set also transcribed training data, and a single linguistic supervisor settled the language questions of the whole project.
- **Unversioned normalization.** The normalization script has no version number yet.
- **Reproducibility.** The data are not public at the time of writing, so the audit cannot yet be reproduced outside the project.

5.2 Considerations for users

Tonal noise is a property to measure, not to hide. The audit shows that tone is both the most valuable and the least consistent layer of the transcriptions. A benchmark that ignored this would reward systems for imitating annotators. We therefore separate the roles of the data: noisy tone in training is part of the task, while the test set must follow one written convention. The planned double annotation will give the human ceiling against which tone error rates can be read.

Scale and annotation serve different purposes. The untranscribed majority of the corpus is useful on its own for adapting self-supervised encoders. The tone-marked transcriptions make it possible to compare tonal modeling strategies on one language with enough data to separate them. The unreviewed transcriptions offer a realistic setting for learning from unreliable labels, which is the normal condition of low-resource annotation.

Three ways to model tone. We expect users to try three: writing tone in the output vocabulary, with a recognizer that predicts base letters and combining tone marks directly; adding an auxiliary tonal head on the encoder, trained jointly with a tone-free recognizer; or restoring tone marks after recognition, from text alone or from text and acoustic tone predictions. A text-only restorer applied to the tone-free references gives the share of tone that is predictable without audio. That figure is the floor that any acoustic tonal model must beat, and it should be reported with every tonal result.

Guided speech has a cost for evaluation. Guided description made large-scale collection of natural speech feasible, as in other recent African corpora. It also produces repeated sentence frames, and scores on a naive split overstate generalisation. Any future split of this corpus should keep the near-duplicate filtering of Section 4.7.

Recognition data are not synthesis data. The corpus tolerates variability that speech synthesis does not: free phrasing, uneven tone marking and missing punctuation. Using it for synthesis requires a dedicated sub-corpus with few speakers, strict verbatim transcription and double annotation of tone, which is planned as separate work.

Services that handle numbers and names need their own tests. Language-model decoding can replace uncertain acoustic output with fluent but wrong text. Anyone building services around numbers, names or amounts should report entity-level accuracy next to word error rate.

The method may transfer within the Gbe cluster. The normalization rules, the tone-aware metrics and the audit procedure depend on properties that Fongbe shares with related languages such as Ewe and Gen.

5.3 Ethical considerations

All speakers and annotators took part voluntarily and gave informed consent to the recording, transcription and research use of their speech.

Voice is personal data. Speakers are identified only by pseudonymous identifiers, and metadata are limited to coarse categories. Released speech could still be misused, for example for voice imitation, and the licence and data card will state the permitted uses.

Recognizers trained on this corpus will still make errors on tone, numbers and names. They should not be used without human confirmation in emergency, payment or health services.

Finally, an annotation protocol for a language whose written norms are still in flux is a form of language planning. The conventions used here were set for engineering reasons and are submitted to linguists and speakers for revision.

6 Conclusion

We presented a Fongbe speech dataset of about 700 hours, with tone-marked transcriptions for about 351 hours, of which 159 are validated by a second person. It is, to our knowledge, the largest speech resource for the language and the first built to study its tone at scale. We described how it was collected, transcribed and normalized, and we measured the consistency of its tonal annotation instead of assuming it.

The audit found that, before review, a quarter of word forms occurred with several tonal spellings and annotators differed systematically in how densely they marked tone. Review narrowed the gap without closing it, and more than a third of the unreviewed utterances changed. These findings shaped the protocol we propose for using the data: splits disjoint on the speaker identifier and deduplicated, a test set to be harmonised under one written convention, and metrics that score tone explicitly.

Three steps follow. The first is the public release of the data with its convention note, datasheet and scoring script. The second is Fon-Eval, a benchmark built on the harmonised test set once it exists, with the baselines that this paper leaves out. The third is a shared competition on frugal tonal recognition, intended to bring the community to work on tone within pre-trained speech models.

6.1 Acknowledgements

This work was carried out within *J'aime Ma Langue*, a national programme for Beninese languages run by the Institute for Inclusive Digital Africa (IIDiA), through its Regional Lab, in partnership with the Beninese Agency for Information Systems and Digital Affairs (ASIN) and

under the authority of the Ministry of Digital Affairs of Benin. We thank the Gates Foundation for its financial support. We also thank the 2,420 contributors who recorded their voices, the annotators and 27 validators who transcribed and checked the recordings, and the linguists who reviewed the transcription conventions. We thank ASIN for its support in running the collection platform. Finally, we thank Digital Umuganda for providing the collection platform, and ASIN for its support in running it. The views expressed are those of the authors.

References

Mahounan Pericles Adjovi, Victor Olufemi, Roald Eiselen, and Prasenjit Mitra. From speech to text corpora: Evaluating ASR-based data acquisition for low-resource Fongbe and Hausa. arXiv:2606.22274, 2026. URL https://arxiv.org/abs/2606.22274.

Benjamin Akera, Evelyn Nafula, Patrick Walukagga, Gilbert Yiga, John Quinn, and Ernest Mwebaze. How much speech data is necessary for ASR in African languages? an evaluation of data scaling in Kinyarwanda and Kikuyu. arXiv:2510.07221, 2025. URL https://arxiv.org/abs/2510.07221.

Jesujoba O. Alabi, Xuechen Liu, Dietrich Klakow, and Junichi Yamagishi. AfriHuBERT: A self-supervised speech representation model for African languages. In *Proceedings of Interspeech 2025*, 2025. arXiv:2409.20201.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common Voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, 2020.

Arun Babu, Changhan Wang, Andros Tjandra, et al. XLS-R: Self-supervised cross-lingual speech representation learning at scale. In *Proceedings of Interspeech 2022*, 2022.

Etienne Barnard, Marelie H. Davel, Charl van Heerden, Febe de Wet, and Jaco Badenhorst. The NCHLT speech corpus of the South African languages. In *Proceedings of the 4th Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU)*, pages 194–200, 2014.

Kaushal Santosh Bhogale, Tahir Javed, Greeshma Susan John, Dhruv Rathi, Akshayasree Padmanaban, Niharika Parasa, and Mitesh M. Khapra. Towards orthographically-informed evaluation of speech recognition systems for Indian languages. arXiv:2603.00941, 2026. URL https://arxiv.org/abs/2603.00941.

Steven Bird. When marking tone reduces fluency: An orthography experiment in Cameroon. *Language and Speech*, 42(1):83–115, 1999a.

Steven Bird. Strategies for representing tone in African writing systems. *Written Language and Literacy*, 2(1):1–44, 1999b.

Marcely Zanon Boito, Vivek Iyer, Nikolaos Lagos, Laurent Besacier, and Ioan Calapodescu. mHuBERT-147: A compact multilingual HuBERT model. In *Proceedings of Interspeech 2024*, 2024.

Antoine Caubrière and Elodie Gauthier. Africa-centric self-supervised pre-training for multilingual speech representation in a sub-Saharan context. In *Proceedings of the 5th Workshop on African Natural Language Processing (AfricaNLP 2024)*, 2024.

Dazhong Chen, Yi-Cheng Lin, Yuchen Huang, Ziwei Gong, Di Jiang, Zeying Xie, and Yi R. Fung. CantoASR: Prosody-aware ASR-LALM collaboration for low-resource Cantonese. arXiv:2511.04139, 2025. URL https://arxiv.org/abs/2511.04139.

Fei-Yueh Chen, Lateef Adeleke, and C. M. Downey. Linguistically informed evaluation of multilingual ASR for African languages. arXiv:2602.04716, 2026. URL https://arxiv.org/abs/2602.04716.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. FLEURS: Few-shot learning evaluation of universal representations of speech. In *Proceedings of the 2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, 2022.

Abdoulaye Diack et al. WAXAL: A large-scale multilingual African language speech corpus. arXiv:2602.02734, 2026. URL https://arxiv.org/abs/2602.02734.

Yacouba Diarra, Panga Azazia Kamate, Nouhoum Souleymane Coulibaly, and Michael Leventhal. Kunnafonidilaw ka cadeau: An ASR dataset of present-day Bambara. arXiv:2512.19400, 2025. URL https://arxiv.org/abs/2512.19400.

Bonaventure F. P. Dossou and Chris C. Emezue. FFR v1.1: Fon-French neural machine translation. arXiv preprint, 2020.

Bonaventure F. P. Dossou and Chris C. Emezue. OkwuGbé: End-to-end speech recognition for Fon and Igbo. arXiv:2103.07762, 2021. URL https://arxiv.org/abs/2103.07762.

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

Aparna Elangovan, Jiayuan He, and Karin Verspoor. Memorization vs. generalization: Quantifying data leakage in NLP performance evaluation. In *Proceedings of EACL 2021*, pages 1325–1335, 2021.

AbdelRahim Elmadany, Sang Yun Kwon, Hawau Olamide Toyin, Alcides Alcoba Inciarte, Hanan Aldarmaki, and Muhammad Abdul-Mageed. Voice of a continent: Mapping Africa's speech technology frontier. arXiv:2505.18436, 2025. URL https://arxiv.org/abs/2505.18436.

Chris Emezue, The NaijaVoices Community, Busayo Awobade, Abraham Toluwase Owodunni, Handel Emezue, Gloria Monica Tobechukwu Emezue, Nefertiti Nneoma Emezue, Sewade Ogun, Bunmi Akinremi, David Ifeoluwa Adelani, and Chris Pal. The NaijaVoices dataset: Cultivating large-scale, high-quality, culturally-rich speech data for African languages. In *Proceedings of Interspeech 2025*, 2025. arXiv:2505.20564.

Dongji Gao, Hainan Xu, Desh Raj, Leibny Paola Garcia Perera, Daniel Povey, and Sanjeev Khudanpur. Learning from flawed data: Weakly supervised automatic speech recognition. arXiv:2309.15796, 2023. URL https://arxiv.org/abs/2309.15796.

Elodie Gauthier, Laurent Besacier, Sylvie Voisin, Michael Melese, and Uriel Pascal Elingui. Collecting resources in sub-Saharan African languages for automatic speech recognition: A case study of Wolof. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, 2016.

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

Mor Geva, Yoav Goldberg, and Jonathan Berant. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of EMNLP-IJCNLP 2019*, pages 1161–1166, 2019.

Pegah Ghahremani, Bagher BabaAli, Daniel Povey, Korbinian Riedhammer, Jan Trmal, and Sanjeev Khudanpur. A pitch extraction algorithm tuned for automatic speech recognition. In *Proceedings of ICASSP 2014*, 2014.

Alexander Gutkin, Işın Demirşahin, Oddur Kjartansson, Clara Rivera, and Kólá Túbòsún. Developing an open-source corpus of Yoruba speech. In *Proceedings of Interspeech 2020*, pages 404–408, 2020.

David Harwath and James Glass. Deep multimodal semantic embeddings for speech and images. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, 2021.

Sukairaj Hafiz Imam, Tadesse Destaw Belay, Kedir Yassin Husse, Ibrahim Said Ahmad, Idris Abdulmumin, Hadiza Ali Umar, Muhammad Yahuza Bello, Joyce Nakatumba-Nabende, Seid Muhie Yimam, and Shamsuddeen Hassan Muhammad. Automatic speech recognition (ASR) for African low-resource languages: A systematic literature review. arXiv:2510.01145, 2025a.

Sukairaj Hafiz Imam, Babangida Sani, Dawit Ketema Gete, Bedru Yimam Ahamed, Ibrahim Said Ahmad, Idris Abdulmumin, Seid Muhie Yimam, Muhammad Yahuza Bello, and Shamsuddeen Hassan Muhammad. Automatic speech recognition for African low-resource languages: Challenges and future directions. In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, pages 89–94, 2025b.

IWSLT. IWSLT 2025 low-resource speech translation track: Data and task description. https://iwslt.org/2025/low-resource, 2025.

Nanling Jiang et al. Latent softmax for data-efficient phoneme-based multilingual ASR across tonal and non-tonal languages. arXiv:2608.01281, 2026. URL https://arxiv.org/abs/2608.01281.

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

D. Fortuné Kponou, Fréjus A. A. Laleye, and Eugène Cokou Ezin. FFSTC: Fongbe to French speech translation corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7270–7276, Torino, Italia, 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.638.

D. Fortuné Kponou, Salima Mdhaffar, Fréjus A. A. Laleye, Eugène C. Ezin, and Yannick Estève. Extending the Fongbe to French speech translation corpus: Resources, models and benchmark. In *Proceedings of Interspeech 2025*, 2025a. URL https://www.isca-archive.org/interspeech_2025/kponou25_interspeech.pdf.

Fortuné Kponou, Fethi Bougares, Yannick Estève, and Salima Mdhaffar. LIA and ELYADATA systems for the IWSLT 2025 low-resource speech translation shared task. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, 2025b.

Fortuné Kponou, Salima Mdhaffar, Fréjus A. A. Laleye, Eugène Cokou Ezin, and Yannick Estève. FFSTC 2: Extending the Fongbe to French speech translation corpus. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, 2025c.

Fréjus A. A. Laleye. Fongbe speech dataset. Zenodo, version 1.0, CC BY 4.0, 2022. URL https://doi.org/10.5281/zenodo.6604637.

Fréjus A. A. Laleye, Laurent Besacier, Eugène C. Ezin, and Cina Motamed. First automatic Fongbe continuous speech recognition system: Development of acoustic models and language models. In *Proceedings of the 2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*, 2016. URL https://annals-csis.org/Volume_8/drp/153.html.

Mingfei Lau, Q. Chen, Yeming Fang, Tingting Xu, T. Chen, and Pavel Golik. Data quality issues in multilingual speech datasets: The need for sociolinguistic awareness and proactive language planning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)*, 2025. URL https://aclanthology.org/2025.acl-long.370/.

Claire Lefebvre and Anne-Marie Brousseau. *A Grammar of Fongbe*. Mouton de Gruyter, Berlin, 2002.

Siyu Liang and Gina-Anne Levow. Breaking the transcription bottleneck: Fine-tuning ASR models for extremely low-resource fieldwork languages. arXiv:2506.17459, 2025. URL https://arxiv.org/abs/2506.17459.

Zoey Liu, Justin Spence, and Emily Prud'hommeaux. Investigating data partitioning strategies for crosslinguistic low-resource ASR evaluation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2023)*, pages 123–131, 2023. URL https://aclanthology.org/2023.eacl-main.10.

Loren Lugosch and Vikrant Singh Tomar. Tone recognition using lifters and CTC. In *Proceedings of Interspeech 2018*, 2018.

Ian Maddieson. Tone. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*, chapter 13. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. URL https://wals.info/chapter/13.

Quinten McNamara, Miguel Ángel del Río Fernández, Nishchal Bhandari, Martin Ratajczak, Danny Chen, Corey Miller, and Migüel Jetté. Style-agnostic evaluation of ASR using multiple reference transcripts. arXiv:2412.07937, 2024. URL https://arxiv.org/abs/2412.07937.

Josh Meyer, David Ifeoluwa Adelani, Edresson Casanova, et al. BibleTTS: A large, high-fidelity, multilingual, and uniquely African speech corpus. In *Proceedings of Interspeech 2022*, 2022.

Dena Mujtaba, Nihar R. Mahapatra, Megan Arney, J. Scott Yaruss, Hope Gerlach-Houck, Caryn Herring, and Jia Bin. Lost in transcription: Identifying and quantifying the accuracy biases of automatic speech recognition systems against disfluent speech. arXiv:2405.06150, 2024.

Alvin Nahabwe, Sulaiman Kagumire, Denis Musinguzi, Bruno Beijuka, Jonah Mubuuke Kyagaba, Peter Nabende, Andrew Katumba, and Joyce Nakatumba-Nabende. Benchmarking automatic speech recognition models for African languages. arXiv:2512.10968, 2025. URL https://arxiv.org/abs/2512.10968.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, et al. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020.

Tolúlọpẹ́ Ógúnrèmí, Kọ́lá Túbọ̀sún, Anuoluwapo Aremu, Iroro Orife, and David Ifeoluwa Adelani. ÌròyìnSpeech: A multi-purpose Yorùbá speech corpus. arXiv:2307.16071, 2023. URL https://arxiv.org/abs/2307.16071.

Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, et al. AfriSpeech-200: Pan-African accented speech dataset for clinical and general domain ASR. *Transactions of the Association for Computational Linguistics*, 11, 2023.

Joy Olusanya. Tone in Yoruba ASR: Evaluating the impact of tone recognition on transformer-based ASR models. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026)*, pages 149–156, 2026. URL https://aclanthology.org/2026.loreslm-1.14.

Omnilingual ASR Team. Omnilingual ASR: Open-source multilingual speech recognition for 1600+ languages. arXiv preprint, 2025.

Iroro Orife. Attentive sequence-to-sequence learning for diacritic restoration of Yorùbá language text. In *Proceedings of Interspeech 2018*, 2018.

Daniel S. Park, Yu Zhang, Ye Jia, Wei Han, Chung-Cheng Chiu, Bo Li, Yonghui Wu, and Quoc V. Le. Improved noisy student training for automatic speech recognition. In *Proceedings of Interspeech 2020*, 2020.

John F. Pitrelli, Mary E. Beckman, and Julia Hirschberg. Evaluation of prosodic transcription labeling reliability in the ToBI framework. In *Proceedings of the 3rd International Conference on Spoken Language Processing (ICSLP 1994)*, pages 123–126, 1994.

Barbara Plank. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of EMNLP 2022*, pages 10671–10682, 2022.

Vineel Pratap, Awni Hannun, Gabriel Synnaeve, and Ronan Collobert. Star temporal classification: Sequence classification with partially labeled data. arXiv:2201.12208, 2022. URL https://arxiv.org/abs/2201.12208.

Vineel Pratap, Andros Tjandra, Bowen Shi, et al. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25, 2024.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*, 2023.

Seamless Communication. SeamlessM4T: Massively multilingual and multimodal machine translation. arXiv:2308.11596, 2023.

Gaofei Shen, Michaela Watkins, Afra Alishahi, Arianna Bisazza, and Grzegorz Chrupała. Encoding of lexical tone in self-supervised models of spoken language. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2024)*, 2024. URL https://aclanthology.org/2024.naacl-long.239.

Claytone Sikasote, Kalinda Siaminwe, Stanly Mwape, Bangiwe Zulu, Mofya Phiri, Martin Phiri, David Zulu, Mayumbo Nyirenda, and Antonios Anastasopoulos. Zambezi Voice: A multilingual speech corpus for Zambian languages. In *Proceedings of Interspeech 2023*, pages 3984–3988, 2023.

Andreas Stolcke and Jasha Droppo. Comparing human and machine errors in conversational speech transcription. In *Proceedings of Interspeech 2017*, 2017.

Tianyi Xu, Xuan Ouyang, Binwei Yao, Shoua Xiong, Sara Misurelli, Maichou Lor, and Junjie Hu. SITA: Learning speaker-invariant and tone-aware speech representations for low-resource tonal languages. arXiv:2601.09050, 2026. URL https://arxiv.org/abs/2601.09050.