Dealing with the Hard Facts of Low-Resource African NLP

Yacouba Diarra, Nouhoum Souleymane Coulibaly, Panga Azazia Kamaté, Madani Amadou Tall, Emmanuel Élisé Koné, Aymane Dembélé, Michael Leventhal
RobotsMali AI4D Lab, Bamako, Mali; **Correspondence:** research@robotsmali.org

Abstract

Creating speech datasets, models, and evaluation frameworks for low-resource languages remains challenging given the lack of a broad base of pertinent experience to draw from. This paper reports on the field collection of 612 hours of spontaneous speech in Bambara, a low-resource West African language; the semi-automated annotation of that dataset with transcriptions; the creation of several monolingual ultra-compact and small models using the dataset; and the automatic and human evaluation of their output. We offer practical suggestions for data collection protocols, annotation, and model design, as well as evidence for the importance of performing human evaluation. In addition to the main dataset, multiple evaluation datasets, models, and code are made publicly available.

1 Introduction

End-to-end ASR (E2E-ASR) systems for languages with large amounts of text data, especially English, have achieved human-level performance on several benchmarks (Xiong et al., 2016). In contrast, training E2E-ASR for low-resource languages remains challenging due to the considerable amounts of labeled data and computational resources required by modern deep learning architectures (Kaplan et al., 2020).

Until recently, African language aligned data for ASR existed only as a minute portion of large multilingual datasets, often primarily for benchmarking purposes (Ardila et al., 2020; Goyal et al., 2022), rather than for training models intended to be deployed in systems that recognize real-world speech. No open ASR model exists for the vast majority of the 2000+ languages on the continent.

In pursuit of what they call *omnilingualism*, Meta released the Massively Multilingual Speech (Pratap et al., 2023) and Omnilingual ASR (Omnilingual ASR Team et al., 2025) model suites in 2023 and 2025. Their approach used massive self-supervised learning (Baevski et al., 2020) and finetuning on small labeled datasets, consisting, principally, of publicly available readings of religious texts in the 2023 release, and data obtained from community-centered crowdsourced data in the 2025 release.[1] This project provided some level of ASR capability for many African and non-African languages for the first time. While a positive development, there is less than 50 hours of data for many of those languages, some with less than 10 hours, a very small fraction of the **120, 710** hours on which the supervised finetuning (SFT) models were trained. The underlying self-supervised encoder (a 7B-parameter wav2vec 2.0 model) was trained on approximately 4.3M hours of unlabeled audio (Omnilingual ASR Team et al., 2025).

For Bambara, a Manding language spoken in several West African countries (primarily in Mali), with more than 15 million L1 and L2 speakers and mutual intelligibility with Malinke, Dioula, and Mandinka, which are spoken by an additional 25 million people, the development of speech recognition technology could affect a population of roughly 40 million (Eberhard et al., 2023). However, as a *low-literacy, predominantly oral language*, Bambara transcription is a hard problem: few speakers can write it, and even those who can lack the facility to do so quickly and easily (Diarra et al., 2025b).

The CMU Wilderness Multilingual Speech Dataset, a dataset of aligned sentences and audio for some 700 languages based on readings of the New Testament, is, to the best of our knowledge, the first mention of Bambara in speech corpora

[1] Meta released part of their labeled corpus openly, offering spontaneous speech recordings and their transcriptions for 348 under-served languages, along with training script configuration and docs: https://github.com/facebookresearch/omnilingual-asr/

Proceedings of the 7th Workshop on African Natural Language Processing (AfricaNLP 2026), pages 1–10

prepared specifically to train speech synthesis models (Black, 2019). The dataset was never released as an open resource. Jeli-ASR, a corpus of 30 hours of griot narrations with their transcriptions and French translations, has so far been the only open ASR dataset for Bambara (Diarra et al., 2022). Since its release in 2022, Jeli-ASR has given rise to derivative datasets and has supported the development of the first openly released ASR models for Bambara on Hugging Face[2].

The African Next Voices project (ANV), undertaken by a network of African universities and organizations, recently released what is thought to be the largest dataset of African languages for AI so far (Marivate et al., 2025), with more still to be published. The project aims to record and transcribe over 9,000 hours of speech in 18 languages across South Africa, Kenya, Nigeria, and Mali (after Bambara completed the list as a later addition).

In this paper, we present the Bambara portion of this initiative for which we have collected and annotated 612 hours of spontaneous Bambara speech collected across the southern part of Mali. We share statistics and metadata about the dataset, the collection process and the results of our ASR experiments with models finetuned and tested on a subset of 101 hours.

2 Data Collection and Annotation

In the audio recording phase, we followed an approach similar to Emezue et al., using *facilitators*: individuals with knowledge of the language—L1 or L2 speakers—whom we trained in data collection guidelines and in the use of our mobile data collection app.[3] The guidelines covered requirements for the recording environment, quality checks for background noise, and management of participants' contributions with respect to voice quality, staying on topic, and minimizing code-switching to French. Pronunciation, often an issue when recordings are based on read speech, was rarely a concern here, as we recorded spontaneous speech on familiar topics from L1/L2 contributors. In total, we collected 626.32 hours of audio and processed 612 hours to create the dataset.

[2]oza75 released a finetuned Whisper model in early 2024 (later taken down), followed by several releases by RobotsMali in early 2025

[3]We also open-source this app, a minimalist Flutter-based tool designed with a simple user interface to minimize user training time: https://github.com/RobotsMali-AI/Africa-Voice-App

The raw recordings were then segmented using Silero VAD's open voice activity detection model (Silero Team, 2024), retaining on average 70% of the original duration and yielding 423 hours of speech chunks ranging from 240 milliseconds to 30 seconds. This step also removed long silences and inaudible speech from the recordings, increasing the amount of usable speech for the transcription pipeline and eliminating manual segmentation (Li et al., 2019). Almost all segments are mono-speaker, although a small number of recordings include brief facilitator speech; overlaps are rare. The segments were first pre-transcribed with RobotsMali/soloni-114m-tdt-ctc-v0, and human transcribers were tasked with reviewing and correcting these model-generated transcriptions rather than transcribing from scratch. We then finetuned RobotsMali/soloni-114m-tdt-ctc-v2 on 98 hours of human-corrected transcripts and re-transcribed the segments with this model. WER and CER metrics were used to compare the two sets of transcriptions, and human review and correction continued on the new model outputs for a period of time to assess their impact on the annotation process (Section 5.2). Further details on transcription guidelines and the labeling interface are provided in Appendices A and B.

3 The ANV Bambara Dataset

The African Next Voices Dataset is the largest open Bambara ASR corpus collected to date. Comprised of natural, spontaneous speech, often from people with great knowledge in domains such as health, agriculture, the food industry, art, and more, we have captured a profound and authentic snapshot of the Malian society and culture in relatively pure Bambara, with the habitual code-switching to French minimized. The dataset collection was also designed to limit the variance of accents and speaking styles, focusing on the southern regions of Mali, relatively close to the capital. While a truly representative dataset would include code-switching and regional accent variation, the objective fixed for ANV was to gather homogeneous data that would simplify training and provide a baseline for a group of low-resource languages with limited NLP support.

3.1 Profile of the Dataset

The dataset features 512 unique speakers from Bamako, the capital, and four localities between

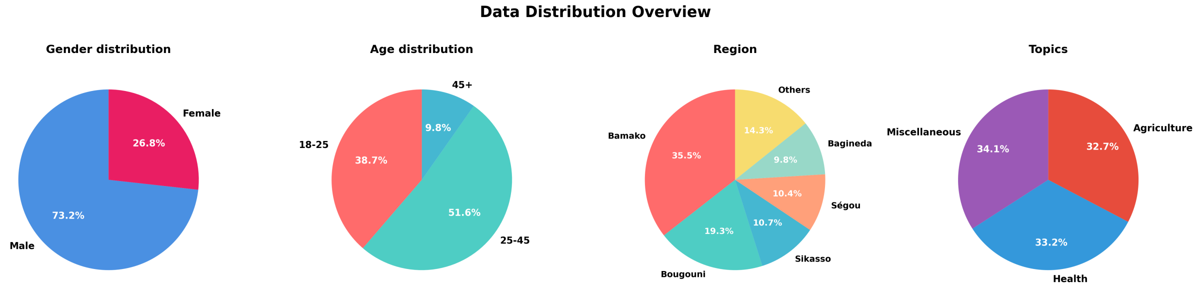

Figure 1: Statistics overview charts of the African Next Voices Bambara dataset: Age, Gender, Region and topics distribution. The first three charts are calculated with respect to the number of speakers while the topics distributions are expressed in durations. The locations represented as 'others' refer to rural areas/villages around the 5 other main regions

25 and 300 km from Bamako: Ségou, Sikasso, Bagineda, and Bougouni. Figure 1 presents an overview of the gender, age, region, and subject-matter distributions in the dataset. The "Miscellaneous" category encompasses topics ranging from education to social norms and beliefs, history, industry, art, and fashion. The speaker distribution is not gender-balanced, reflecting cultural and security constraints at the time of collection.

The average segment duration is $\approx$ 2 seconds. Chunks shorter than one second were not uploaded in the annotation pipeline as they often consist of formulaic expressions, discourse markers and set responses of 1 to 3 words such as nse, nba (female or male response to a salutation), ayiwa (a term used to express agreement or closure) or nka (but). These short segments are transcribed accurately by the models at much lower error rates than longer, novel utterances (Tall, 2025). They are a significant percentage of the total dataset such that including them would skew WER/CER measurements while not contributing to ASR performance. The Hugging Face dataset is divided into 3 subsets totaling $\mathbf{874,762}$ utterances totaling approximately 423 hours. Each subset contains the audio segments and two sets of accompanying transcriptions labeled either **v1**, created by soloni-v0, trained mainly on Jeli-asr, and **v2** created by soloni-v2, finetuned from soloni-v0 using 98 hours of segments and human-corrected transcriptions collected over the course of the project.

- **The 'human-corrected' subset**: A 159 hour subset ($260,008$ utterances) with human reviewed, corrected and validated transcriptions. This subset is the only one with a 'text' attribute containing a transcription that has gone through human review and correction.
- **The 'model-annotated' subset**: A 212 hour subset ($355,571$ utterances) that has model-generated transcriptions that have not been reviewed by humans. This subset has only **v1** and **v2** labels corresponding to the model used to generate transcriptions.
- **The 'short' subset**: A 52 hour subset ($259,183$ utterances) of duration less than one second that we have filtered out from the pool of segments to be annotated. Those short utterances are model-annotated and have **v1** and **v2** labels.

We have also released the original 612 hours dataset comprised of 1777 raw recordings ranging from 8 seconds to 1.48 hours, with all the associated segment timestamps, the anonymized metadata, the SNR quality check results and all the preprocessing code[4].

3.2 Signal-to-Noise Ratio as a proxy to Audio Quality

We use a Voice Activity Detection (VAD) based method to estimate SNR. VAD output is used to separate the signal into two distinct regions: speech activity ($\text{vad}[n] = 1$) and voice-inactive ($\text{vad}[n] = 0$). We use VAD to estimate the speech and noise power instead of a histogram-based approach such as the standard NIST SNR method. In our setup, SNR is defined, using VAD, as the ratio between the estimated speech power (from speech activity regions) and the estimated noise power (typically the average power in silence regions) from the same recording (Vondrasek and Pollák, 2005;

[4]We have made the recordings and metadata available through Google Cloud Storage. The link will be found in the GitHub repository holding the code: https://github.com/RobotsMali-AI/afvoices

Silero Team, 2024.) Figure 2 shows the distribution of SNR values from the unsegmented recordings in the dataset.

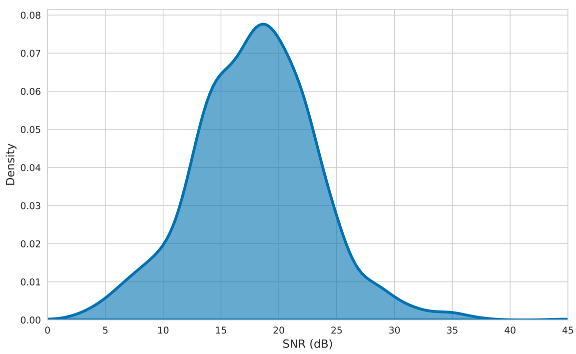

Figure 2: Density Distribution of Signal-to-Noise Ratio values in the African Next Voices Bambara Dataset. Note that the SNR values are not bounded.

The classical SNR definition yields a different value distribution than VAD-based estimates, so we follow Xu et al., who report VAD-based SNRs for several noise types and levels at 5, 10, and 20 dB and treat 30 dB as clean speech; Table 1 groups our recordings into five bands using thresholds derived from their datapoints. We prefer a VAD-based SNR here because ASR performance depends on the quality of speech frames rather than the entire waveform, including silence. Estimating signal power over VAD-identified speech regions and noise power over non-speech regions yields an SNR measure that better reflects recognition difficulty in our small corpus and allows us to retain and prioritize high-quality speech segments from otherwise noisy files.

SNR Category	Threshold (dB)	Recordings
Very Low SNR	< 0	1
Low SNR	$[0, 5)$	15
Medium SNR	$[5, 15)$	486
High SNR	$[15, 25)$	1135
Very High SNR	≥ 25	140
	Total Audios	**1777**

Table 1: Distribution of Audio Recordings by Signal-to-Noise Ratio (SNR) Category.

71.75% of the recordings fall into the 'High SNR' to 'Very High SNR' categories; this indicates that the dataset consists of relatively clean audio recordings.

4 ASR Experiments

We performed experiments with a subset of our human-corrected transcribed segments to explore the potential of the ANV Bambara dataset for monolingual ASR modeling. We finetuned the models from our earlier experiments with Jeli-asr —themselves finetuned from different models of NVIDIA's Parakeet family of English-trained ASR models and on QuartzNet— and evaluate all the models on both the test set of our experiment (Afvoices Test) and a smaller, more heterogeneous benchmark (Nyana Eval) that we also introduce in this paper. We also report on human evaluation by native speakers, comparing results obtained from the latest finetuned models and their predecessors on the Nyana Eval benchmark (section 5.2).

Experimental setup: For our experiments, we finetuned open-source models based on NVIDIA's Parakeet family and QuartzNet.

We finetuned a 114M- and two 600M-parameter Parakeet models[5], but as we did not perform human evaluation on the larger model, we only report on the 114M model, `soloni-114m-tdt-ctc-v0`, in this study. `soloni-114m-tdt-ctc-v0` uses a Fast-Conformer encoder (Rekesh et al., 2023) and a hybrid decoding setup with two independent but jointly trained decoders: a Token-and-Duration Transducer (TDT) decoder —an extension of the RNN-Transducer that predicts both a token and its duration (Xu et al., 2023)—and a convolutional decoder trained with a Connectionist Temporal Classification (CTC) loss (Graves et al., 2006). This dual-decoder design makes the model particularly interesting for analyzing how the two decoding approaches behave under the same training conditions. `soloni-114m-tdt-ctc-v0` also provides insight into how the architecture will perform when scaling up to the larger models in the family.

`stt-bm-quartznet15x5-v0` is a finetune of NVIDIA's 18M parameter, ultra-compact QuartzNet model, an end-to-end convolutional architecture with 1D time-channel separable convolutions (Kriman et al., 2019). This model addresses a particularly critical use case in Mali where a large portion of the population does not have access to internet connectivity. We had al-

[5] We trained the Parkeet 600M model with an auto regressive decoder and 600M model with a convolutional decoder. While these models are not further discussed in this paper, all our models with their associated metrics can be found in RobotsMali's Hugging Face repository.

ready deployed this model in a Bambara-language ASR-based reading tutor app[6] that runs locally in low-end smartphones.

QuartzNet is a character-based decoding model. The vocabulary for the dataset consists of 38 unique characters, including the 30 letters of the Bambara alphabet, 5 accented French vowels, whitespace, hyphen (used in some compound words) and apostrophe (largely used in writing contractions). For soloni we train a SentencePiece tokenizer with a vocab size of 512. We had 4 NVIDIA A100 GPUs with a combined 320GB of VRAM for the experiment.

Training Data: We finetuned the two models on 98 hours of voice data with human-corrected transcripts, consisting of $167,816$ utterances, and tested on 3 hours (5175 samples). We implemented and applied most of the normalization steps described by Zupon et al. before training, but we did not remove any of the acoustic event tags (presented in Appendix A) as we wanted to model those events as well.

Training configurations: We first trained soloni for 110k steps on 2 GPUs, with 32 batch size, using the AdamW optimizer and Noam scheduler with learning rate scaling factor of 0.003 and a 10% warmup ratio (Vaswani et al., 2017). Then we trained the resulting model for 100k more steps on all 4 GPUs, this time with an LR scaling factor of 1.5 and a 2% warmup ratio, all with no gradient accumulation and `bf16` precision.

We trained QuartzNet for 65k steps on 4 GPUs, with 64 batch size, using the Novograd optimizer (Ginsburg et al., 2020) and a Cosine LR scheduler with a 1×10^3 and 1×10^6 upper and lower bounds and $6,000$ warmup steps.

5 Evaluation of the Models

We evaluated `soloni-114m-tdt-ctc-v0` both with the CTC and with TDT decoders, and the QuartzNet model on the Afvoices test set and a smaller benchmark, ɲɛna (transliterated to Nyana for English keyboards and nyana-eval for identification on Hugging Face) [7] , that we compiled through the stratified sampling of 15 audio files from each of:

- the test set of Kunkado (Diarra et al., 2025a);
- a generally unused subset of *Jeli-asr* (street interviews) that we cleaned beforehand (Diarra et al., 2022);
- crowd-sourced recordings of readings of excerpts of the books from the GAIFE project (Tapo et al., 2025).

5.1 WER evaluation

Table 2 presents the Word Error Rates and Character Error Rates of the two models before and after our finetuning experiments. We show a significant improvement for all models across all metrics and benchmarks, up to **37% WER improvement** for soloni-114m-tdt, the best performing model overall. The relatively smaller WER improvement on the more challenging Nyana Eval benchmark, containing much noisier—sometimes multi-speaker—audio from street interviews and radio recordings than the Afvoices training and test data, highlights the potential limitations of the model in real-world deployment scenarios.

We note that the CTC branch of soloni lost its edge over TDT branch on the Afvoices test set when we increased the amount of training data from the 30 hours of Jeli-asr to 98 hours in this experiment, confirming that, for many sequence modeling tasks, autoregressive architectures outperform non-autoregressive ones when we scale training data (Graves, 2012; Li et al., 2020).

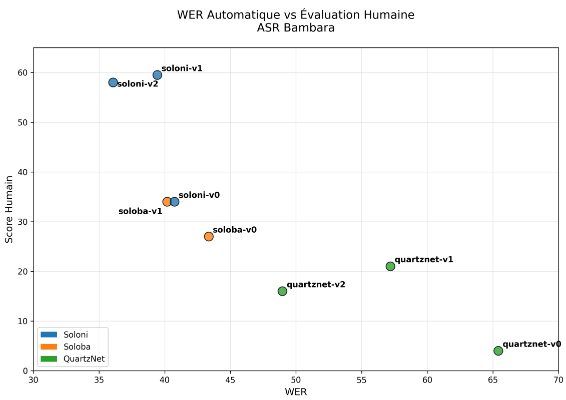

Figure 3: WER vs human evaluation. Figure from (Tall, 2025)

5.2 Human Evaluation

We performed a detailed human analysis of the outputs of several RobotsMali ASR models, including the models from this experiment, using

[6] Our reading tutor app using the QuartzNet model, An Bɛ Kalan, is available for iOS and Android

[7] ɲɛna means “opinion”; this emphasizes the human-evaluation focus of the dataset. We also release it on Hugging Face: https://huggingface.co/datasets/RobotsMali/nyana-eval

Model	WER (%) ↓		CER (%) ↓	
	Afvoices Test	Nyana Eval	Afvoices Test	Nyana Eval
soloni-114m (CTC)				
Unfinetuned (v0)	43.12	40.75	23.48	24.7
Finetuned (v2)	29.05	**36.07**	13.41	**20.04**
soloni-114m (TDT)				
Unfinetuned (v0)	45.52	47.1	26.68	31.27
Finetuned (v2)	**28.58**	38.13	**12.94**	22.3
Quartznet (CTC)				
Unfinetuned (v0)	73.66	65.42	37.85	30.66
Finetuned (v2)	42.57	48.97	18.70	24.22

Table 2: ASR experiment metrics: We apply the same normalization steps to our test sets and this time we remove the acoustic event tags from both the reference and the prediction before calculating the WER and CER. The values in bold highlight the best performances per metric

Nyana Eval as test data (Tall, 2025). Across models, we observed systematic difficulties with disfluencies, proper names, code-switching, and overlapping speech. The highest-rated model in this evaluation is `soloni-v1`, a finetune of `soloni-v0` trained on `RobotsMali/kunkado` (Diarra et al., 2025a), a dataset composed of everyday speech. `soloni-v2`, finetuned on the Afvoices dataset with minimized noise, code-switching, and voice overlap, was judged slightly less robust by human evaluators on the more natural recordings in Nyana Eval, despite achieving a better WER (36.07% vs. 39.44%). Figure 3 plots the WER on Nyana Eval against the corresponding human evaluation scores.

We also report speed gains after replacing the **v0** transcriptions with **v2** in the human review-and-correction pipeline. The transcription team completed 45 hours of audio in approximately 800 hours of work, corresponding to a **17×** real-time factor (17 hours of annotation per hour of speech). Using `soloni-v2` instead of `soloni-v0` yielded a 112% improvement in the rate at which human-corrected transcriptions could be produced, compared to an earlier study in which a 36× ratio was observed (Diarra et al., 2025b).

6 Conclusion

We released 612 hours of spontaneous Bambara speech and a 423-hour segmented corpus, together with metadata, VAD-based SNR estimates, transcription guidelines, a minimalist mobile recording app, multiple evaluation sets, and monolingual ASR models (an ultra-compact QuartzNet variant and a 114M-parameter Parakeet-based model). For a low-literacy, predominantly oral language, this substantially increases the pool of publicly available resources.

Using `soloni-114m-tdt-ctc-v0` in a human–model loop enabled semi-automated annotation, reducing the time required for corrected transcriptions and improving WER/CER on both in-domain and more challenging benchmarks. Human evaluation on Nyana Eval revealed systematic gaps between automatic string-based metrics and native-speaker judgments.

Taken together, these elements define a practical workflow for dealing with the hard facts of low-resource African NLP: targeted field collection with trained facilitators, noise-aware preprocessing, semi-automated annotation, and evaluation protocols that combine automatic and human measures while accounting for deployment on modest hardware. Future work will extend this approach to other Manding languages and to datasets that more fully reflect real-world speech.

Limitations

The African Next Voices (ANV) Bambara dataset was meticulously designed to provide a high-quality, clean foundation for ASR research. While successful in establishing the largest open corpus for the language, this intentional simplification inherently introduces limitations when considering the deployment of derived models in authentic, unconstrained Malian contexts. The core limitation lies in the necessary trade-off between controlled data collection and the complex nature of the real-world environment. We can enumerate the following key limitations:

- **Acoustic Purity vs Real-World challenges:** The emphasis on clean recordings resulted in a dataset where 71.75% of the recordings were classified with 'High SNR' or 'Very High SNR'. This limits the model's exposure to challenging acoustic conditions—such as urban street noise and background voices that characterize typical Malian environments. Consequently, models trained exclusively on this corpus may exhibit a noticeable drop in performance when deployed in the real world.

- **Suppression of Code-Switching and Multilingualism:** The transcription protocol's mandate to replace code-switched terms and foreign words with the generic `[cs]` tag, or to force a Bambara-phonology transcription, simplifies the target vocabulary for ASR training. However, this approach sacrifices linguistic realism. The resulting models are fundamentally unprepared to transcribe the common, fluid shifts between Bambara and French, which are integral to spontaneous speech.

Ultimately, the ANV Bambara dataset represents a simplified version of the language's acoustic and linguistic reality. While this simplification provides a more stable foundation for core ASR research, it comes at the cost of real-world robustness. Practitioners seeking to deploy these models in authentic Malian contexts characterized by inherent noise, fluent code-switching, and diverse accents must anticipate the need for a targeted domain adaptation.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common Voice: A massively-multilingual speech corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 4211–4215, Marseille, France. European Language Resources Association (ELRA).

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Preprint*, arXiv:2006.11477.

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

Sebastien Diarra, Michael Leventhal, and Allahsera Auguste Tapo. 2022. Robotsmali griots speech dataset, and asr. `https://github.com/robotsmali-ai/jeli-asr/`.

Yacouba Diarra, Nouhoum Coulibaly, Panga Azazia Kamaté, and Michael Leventhal. 2025a. kunnafonidilaw ka cadeau: an ASR dataset to power the development of models that understand present-day bambara. Hugging Face Datasets. Arxiv coming soon.

Yacouba Diarra, Nouhoum Souleymane Coulibaly, and Michael Leventhal. 2025b. Cost analysis of human-corrected transcription for predominately oral languages. *Preprint*, arXiv:2510.12781.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2023. *Ethnologue: Languages of the World*, 26th edition. SIL International, Dallas, TX.

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

Boris Ginsburg, Patrice Castonguay, Oleksii Hrinchuk, Oleksii Kuchaiev, Vitaly Lavrukhin, Ryan Leary, Jason Li, Huyen Nguyen, Yang Zhang, and Jonathan M. Cohen. 2020. Stochastic gradient methods with layer-wise adaptive moments for training of deep networks. *Preprint*, arXiv:1905.11286.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. *Preprint*, arXiv:1211.3711.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural 'networks. In *ICML 2006*, volume 2006, pages 369–376.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Mamadou Konta and Valentin Vydrin. 2014. Propositions pour l'orthographe du bamanankan. *Mandenkan*, (52):3–38.

Samuel Kriman, Stanislav Beliaev, Boris Ginsburg, Jocelyn Huang, Oleksii Kuchaiev, Vitaly Lavrukhin, Ryan Leary, Jason Li, and Yang Zhang. 2019. Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions. *Preprint*, arXiv:1910.10261.

Jinyu Li, Yu Wu, Yashesh Gaur, Chengyi Wang, Rui Zhao, and Shujie Liu. 2020. On the comparison of popular end-to-end models for large scale speech recognition. *Preprint*, arXiv:2005.14327.

Xinjian Li, Zhong Zhou, Siddharth Dalmia, Alan W. Black, and Florian Metze. 2019. Santlr: Speech annotation toolkit for low resource languages. *Preprint*, arXiv:1908.01067.

Vukosi Marivate, Ife Adebara, and Lilian Wanzare. 2025. African languages for ai: the project that's gathering a huge new dataset. The Conversation. Accessed: 2025-11-13.

Omnilingual ASR Team, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Kevin Chan, Chierh Cheng, Joe Chuang, Caley Droof, Mark Duppenthaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, Kehan Lyu, and 13 others. 2025. Omnilingual ASR: Open-source multilingual speech recognition for 1600+ languages.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling speech technology to 1,000+ languages. *Preprint*, arXiv:2305.13516.

Dima Rekesh, Nithin Rao Koluguri, Samuel Kriman, Somshubra Majumdar, Vahid Noroozi, He Huang, Oleksii Hrinchuk, Krishna Puvvada, Ankur Kumar, Jagadeesh Balam, and Boris Ginsburg. 2023. Fast conformer with linearly scalable attention for efficient speech recognition. *Preprint*, arXiv:2305.05084.

Silero Team. 2024. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. `https://github.com/snakers4/silero-vad`.

Madani Amadou Tall. 2025. Analyse comparative humaine des modèles asr bambara de robotsmali.

Allahsera Auguste Tapo, Nouhoum Coulibaly, Seydou Diallo, Sebastien Diarra, Christopher M Homan, Mamadou K. Keita, and Michael Leventhal. 2025. GAIfE: Using GenAI to improve literacy in low-resourced settings. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7914–7929, Albuquerque, New Mexico. Association for Computational Linguistics.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Preprint*, arXiv:1706.03762.

Martin Vondrasek and Petr Pollák. 2005. Methods for speech snr estimation: Evaluation tool and analysis of vad dependency. *Radioengineering*, 14.

Valentin Feodosievich Vydrin. 2022. Vers un dictionnaire orthographique bambara. *Mandenkan : Bulletin Semestriel d'Études Linguistiques Mandé*, (68):59–82.

Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Michael Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig. 2016. Achieving human parity in conversational speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, PP.

Hainan Xu, Fei Jia, Somshubra Majumdar, He Huang, Shinji Watanabe, and Boris Ginsburg. 2023. Efficient sequence transduction by jointly predicting tokens and durations. *Preprint*, arXiv:2304.06795.

Haitian Xu, Zheng-Hua Tan, Paul Dalsgaard, and Børge Lindberg. 2005. Robust speech recognition based on noise and snr classification - a multiple-model framework. In *Interspeech 2005*, pages 977–980.

Andrew Zupon, Evan Crew, and Sandy Ritchie. 2021. Text normalization for low-resource languages of africa. *Preprint*, arXiv:2103.15845.

A Transcription Guidelines

The transcription process for this Bambara dataset was implemented as a **First Review** task, where annotators corrected and validated pre-transcribed audio segments. Annotators were instructed to respect the following rules to simplify and standardize transcription and ensure high-quality data for ASR training and evaluation.

Language and Orthography

- Use the *standardized orthography of Bambara* (Konta and Vydrin, 2014). Correct any orthographic errors or non-standard characters (e.g., accented characters are prohibited, conforming to the ordinance of AMALAN[8]).

[8]We have been working closely with the Académie Malienne des Langues (AMALAN) and the Direction Nationale de l'Education Non Formelle et des Langues Nationales (DNENF-LN) prior to this project and we were able to leverage their expertise for all matters pertaining to national language standards for Bambara.

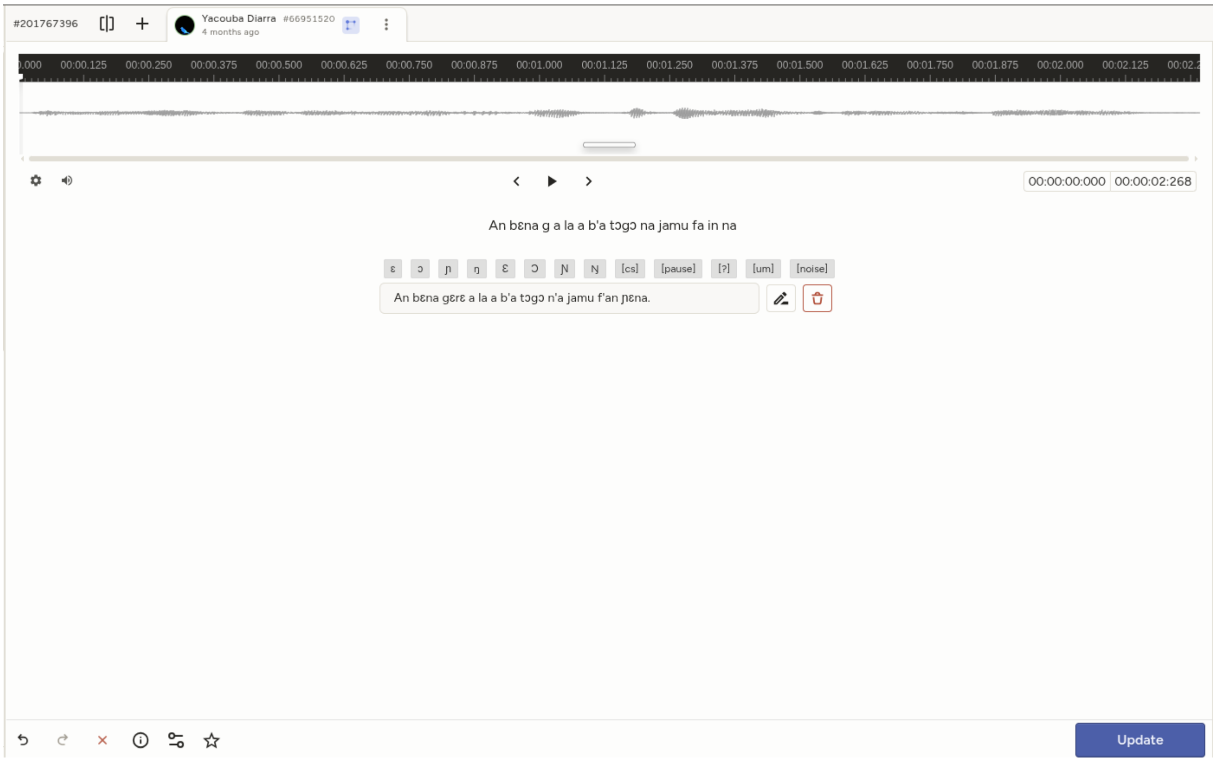

Figure 4: The Labeling Interface for the African Next Voices Bambara Transcription Project. The interface shows the original audio waveform, the automatically generated pre-transcription, and the field for human correction/validation.

- Do not correct the speaker's grammatical errors or pronunciation mistakes.
- Elision and Mispronunciation: Transcribe exactly what is pronounced. For example, transcribe "Ne b'a fɛ" instead of "Ne be a fɛ" if "b'a" was pronounced.
- Proper Nouns: Must be capitalized.
- Repetitions, Disfluencies, and Stuttering: Write out repeated words/phonemes without using ellipses (e.g., "n'i ko ko ko ne kalabanci e dun?").
- Consult the Bamadaba dictionary for language references (Vydrin, 2022).

Numbers, Abbreviations, and Foreign Words

- Numbers: Must be written in full letters exactly as pronounced (e.g., 35 → *bi saba ni duuru / bi saba ni Loru*).
- The Ordinal forms (e.g., *folo*, *filanan*, *sabanan*) must also be written in full letters.
- If a number is pronounced in a foreign language (e.g., French), use the code-switching tag `[cs]`.
- Abbreviations and Acronyms: Transcribe them as pronounced and in *uppercase* (e.g., FIFA, BIM, ORTM). Do not use periods in acronyms.
- Spelled Words: Use hyphens to separate letters (e.g., `K-E-L-E-N`).
- **Code-Switching/Foreign Words:**
  - If the word exists in Bamadaba and was pronounced with the local Bambara phonology, transcribe it as written in the dictionary (e.g., Parce que → *paseke*, passerelle → *pasereli*). This rule exists because Bambara, like many African languages, has borrowed, transformed and standardized many words from the former colonial language.
  - If the word was pronounced with its original foreign pronunciation and/or is not recognized in Bamadaba, replace it with the tag `[cs]` (e.g., `ne ka véhicule` → `ne ka [cs]`).

Acoustic Event Tags (for ASR Modeling)

Acoustic events and background sounds were retained and modeled in the final transcriptions using the following specific tags:

- Inaudible/Incomplete Speech or Overlap: Use the tag `[?]` for incomprehensible, non-audible speech, or speech overlaps.

- Vocalized/Disfluencies (Mouth Sounds): Use the tag `[um]` for sounds like *ɛɛ*, *hum*, *onh*, *ah*, *unhun*, etc.

- Long Silences: Use the tag `[pause]` for silences longer than 5 seconds (or longer than 3 seconds at the beginning or end of a segment). This tag was rarely used after VAD segmentation.

- Background Noise: Use the tag `[noise]` for all occurrences of strong background noise, including applause, coughing, laughter, phone rings, children, etc.

Punctuation

- Standard punctuation (commas, periods, question marks, etc.) should be used.

B Labeling Interface

The data annotation for the ANV Bambara project was performed using a tailored platform built on top of Label Studio (Tkachenko et al., 2020-2022). This interface facilitated the task by presenting pre-transcribed audio segments for human correction and validation. The audio files and their pre-labels were sequentially loaded from Google Cloud Storage into the labeling interface.

As shown in Figure 4, the simple interface provided the following key elements:

- Audio Segment: The interface displays the audio waveform and playback controls, allowing the annotator to listen to the segment.

- Pre-Transcription: The initial transcription was automatically generated by our ASR models.

- Correction Field: The annotator validates and corrects the automatic transcription in a designated field.

- Acoustic Event Tags: A row of buttons provides quick access to the acoustic event tags and the few Bambara characters that are not typically found on a standard keyboard.

This process of pre-transcription followed by human correction minimized human labor for segmentation while optimizing the usable speech data for the transcription pipeline.