Kunnafonidilaw ka Cadeau: an ASR dataset of present-day Bambara

Yacouba Diarra, Panga Azazia Kamaté,
Nouhoum Souleymane Coulibaly, Michael Leventhal
RobotsMali AI4D Lab, Bamako, Mali; **Correspondence:** research@robotsmali.org

Abstract

We present Kunkado, a 160-hour Bambara ASR dataset compiled from Malian radio archives to capture present-day spontaneous speech across a wide range of topics. It includes code-switching, disfluencies, background noise, and overlapping speakers that practical ASR systems encounter in real-world use. We finetuned Parakeet-based models on a 33.47-hour human-reviewed subset and apply pragmatic transcript normalization to reduce variability in number formatting, tags, and code-switching annotations. Evaluated on two real-world test sets, finetuning with Kunkado reduces WER from 44.47% to 37.12% on one and from 36.07% to 32.33% on the other. In human evaluation, the resulting model also outperforms a comparable system with the same architecture trained on 98 hours of cleaner, less realistic speech. We release the data and models to support robust ASR for predominantly-oral languages.

1 Introduction

Data availability for automatic speech recognition (ASR) for the Bambara language has increased significantly this year. For about three years, Jeli-ASR (Diarra et al., 2022), a corpus of 30 hours of transcribed griot narrations, had been the only open ASR dataset for Bambara, but in late 2025, a team at RobotsMali AI4D Lab released a 612 hour dataset as part of the African Next Voices (ANV) project (Diarra et al., 2025a), scaling the amount of open data available by a factor of 20x. In both cases, the audio was recorded during the project in a relatively controlled environment with consistent quality control prior to transcription.

Cost and the challenges of field collection have led to many initiatives aiming to increase speech data for low resource languages (LRLs) to either align data that has already been recorded such as Bible readings (Black, 2019; Pratap et al., 2023) or employ synthetic generation techniques (DeRenzi et al., 2025). The former requires a pretrained acoustic model to produce alignments between acoustic features and the corresponding phonemes in a long transcript (Tsoukala et al., 2023) or an existing aligned set to train a speech synthesizer to generate phones from the transcript and find those phones in the audio (Black, 2019). Synthetic speech data can be useful for robustness training but its value is significantly lower than that of real speech. It also increases the risk of propagating the biases in the generating distribution in the trained models (Rosenberg et al., 2019; Moslem, 2024).

Gender and age are usually the primary concerns with respect to the representativeness of the dataset. Naturalness in spontaneous speech is often neglected even though it is an equally important factor in creating a dataset that truly represents speech in all its dimensions. Many projects curate data following guidelines that hinder the capture of spontaneous speech, seeking to minimize or prohibit code-switching, background noise, slang, and ungrammatical constructions. Such guidelines reduce the usability of the models for real-world deployment scenarios (Diarra et al., 2025a), as these phenomena reflect the realities of daily interactions and the linguistic evolution of many low resource languages. Day-to-day speech in many African languages feature high rates of inter-sentential and intra-sentential language shifts both between African languages and with high-resource colonial languages. Code-switching may enable models to use the high-resource language constantly appearing in sentences and conversations to improve accuracy on the LRL.

Among LRLs, predominantly-oral languages (POL) constitute a large subset where speech, to the almost complete exclusion of writing, has been and remains the dominant means of knowledge transmission. Bambara along with most African languages, is a POL. For many of those languages,

Proceedings of the 7th Workshop on African Natural Language Processing (AfricaNLP 2026), pages 190–196

the only readily-available body of natural communication is radio and television broadcasts. This resource includes background and foreground music, various types of noise, and phone calls in which the audience jumps into the conversation, bringing a variety of accents and dialects (Doumbouya et al., 2021). This abundant source of data is rarely exploited due to the many challenges in transcribing such unpredictable conversations and because it implies renouncing control over topics and audio quality.

In this paper, we introduce **Kunkado**, a 160 hour transcribed ASR dataset compiled from radio archives. The title of this paper, *Kunnafonidilaw ka cadeau* can be translated as "Media's Gift". It is also a good example of the everyday code-switching in Bambara, the word *cadeau* being a direct borrowing from French. The entire dataset was automatically transcribed, with 25% corrected by humans. We report on human evaluation performed on a model trained with the reviewed subset of Kunkado, comparing this result to that of the same model trained with a much larger quantity of curated data. (see Section 4). In the next section, we share insights on handling code-switching and numbers and how trade-offs can be made between consistency and simplicity in transcription.

2 Characteristics of Kunkado

Audio collection and segmentation: We obtained approximately 300 hours of broadcast recordings from 4 Malian radio stations. After segmentation, we only retained segments between 600 milliseconds and 45 seconds of duration. We have released, on Hugging Face[1], 118,925 segments totaling 161.15 hours. Approximately 94% of segments are less than 15 seconds in duration. The mean duration of segments is 4.9 seconds.

For audio segmentation, we employed a simple energy threshold-based method implemented via the `split_on_silence` function from the pydub library (Robert, 2018). This function performs segmentation by analyzing the root mean square (RMS) energy of the audio and splitting the signal where the energy drops below a predefined absolute loudness threshold. Specifically, we used the parameters $min\_silence\_len = 600$ milliseconds and $silence\_thresh = -35$ dBFS. This configuration ensured that the audio stream was only split when the signal's loudness dropped below -35 dB relative to the maximum possible volume for a minimum duration of 600 milliseconds.

While this kind of silence proxy segmentation is much faster than modern voice-activity-detection based methods and does not discard any part of the original recording—i.e., it simply finds endpoints to split on based on the given parameters, yielding $segments\_duration = original\_duration$, it also results in much rougher segmentation and speech cut-offs. We modelled those cut-offs during transcription (see Table 1).

Noise estimation: We calculate signal-to-noise ratio (SNR) as an estimate of the level of noise/non-speech signal in the segmented dataset, using the same implementation and classification thresholds defined by Diarra et al.. **69.2**% of the segments fall above the High SNR category (>15 dB). Although they feature a considerable amount of acoustic non-speech information, this relatively good SNR measurement is explained by the fact that the volume of those events is significantly lower than that of speech, as broadcasts are recorded with professional equipment. The measurements demonstrate that radio data is a quality source to create speech datasets (Doumbouya et al., 2021). Figure 1 shows the density distribution of SNR values in the dataset.

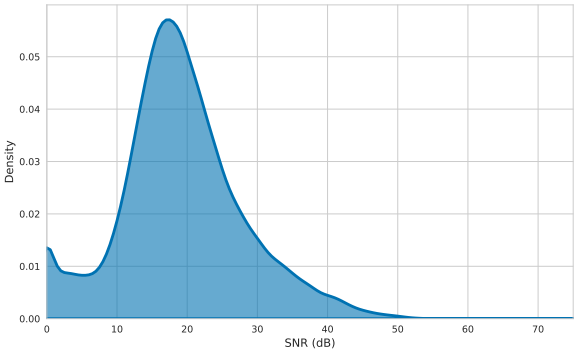

Figure 1: Density Distribution of Signal-to-Noise Ratio values in Kunkado. This figure includes both subsets

Transcription: We have re-engineered transcription, aiming at better matching the requirements of the task to the skill level of annotators that were available to us and speeding up the process. We used the process described in Diarra et al., redefining segment annotation as a first review task where the objective is to correct an automatic transcription generated

[1]The full dataset is released under CC BY-SA 4.0 at RobotsMali/kunkado on HF

Tag	Meaning
<BRUITS>	generic noise
<INCOMPRÉHENSIBLE>	fully inaudible speech
<CHEVAUCHEMENT>	speaker overlap
<RIRES>	laughter
<MUSIQUE>	music / jingle (no lyrics)
<TOUX>	cough
<INVOCATION>	prayers, quranic excerpts
<ECHO>	echo artifact
<APPLAUDISSEMENTS>	applause
<CRIS>	screams
<PLEURES>	crying
__phrase__	double underscores delimit code-switched words and sentences
...	used to mark speech cut-offs and hesitations

Table 1: Tags/Annotations for acoustic and linguistic events captured in Kunkado

by `RobotsMali/soloni-114m-tdt-ctc-v0` Bambara ASR model and to ensure it conforms to our guidelines.

Our guidelines direct annotators to transcribe numbers with Hindu-Arabic numerals, in contrast to orthodox ASR practice where numbers are written out as words. Writing numerals speeds up transcription, reduces potential ambiguity in parsing the quantity, simplifies downstream processing, and is more easily validated by human reviewers. We also have created data for comparison of number formatting approaches in end-to-end ASR—which remains a challenge even in high-resource languages (Huber and Waibel, 2025) Annotators were instructed to use 13 tags to capture as much acoustic and linguistic information as possible. Code-switching to French was written out using the French orthography, ignoring the existence of Bambara-ized spellings commonly, but inconsistently, used for many French words and expressions. Code-switching to Arabic also followed this principle, though transliterated to the Latin alphabet. A transliteration standard was not enforced due to the specialized knowledge required and the complexity that this would have added to the annotation task. While French is often woven into Bambara in a wide variety of ways, Arabic is generally limited to a small set of formulaic Islamic expressions, though these are used frequently. Table 1 presents the complete list of tags and their significations. Other rules on proper nouns, acronyms and spelled-out words are similar to those used in the ANV Bambara project (Diarra et al., 2025a). Although annotators were not required to follow a single standard Bambara orthography, they used the Bamadaba dictionary (Vydrin, 2022) as their primary reference. The annotators used the same data annotation interface described in (Diarra et al., 2025a).

Despite added complexity with respect to tagging and code-switching, our team of seven annotators were able to correct 39.3 hours of segments in roughly 1260 hours of human labor, yielding a 32x ratio, a bit faster than the 36x reference datapoint reported in the transcription cost analysis study by (Diarra et al., 2025b) which used the same model for generating the automated transcriptions. We speculate that more flexible orthography and the use of numerals rather than spelled-out numbers contributed to this 4x difference.

3 ASR Experiments

We finetuned multiple Bambara ASR models, previously trained on Jeli-ASR (Diarra et al., 2022), from RobotsMali's baseline ASR experiments. These models are based on NVIDIA's Parakeet family of monolingual English ASR models (Rekesh et al., 2023). We evaluated all the models on a 5 hour test set taken from the Kunkado data, and Nyana-Eval, a small, stratified human evaluation dataset with only 45 entries of 3 minutes total duration (Diarra et al., 2025a). We report in Section 4 the automatic and human evaluation results for all the RobotsMali soloni models, except the v3 version for which we only report the WER gains since it was not part of the human evaluation[2].

[2]We still release all the other models with the corresponding WER scores. hf.co/RobotsMali/models

Model	WER (%) ↓		CER (%) ↓	
	Kunkado Test	Nyana-Eval	Kunkado Test	Nyana-Eval
soloni (jeli-asr)				
Unfinetuned (v0)	46.91	40.75	30.56	24.71
Finetuned (v1)	39.13	39.44	**20.98**	20.5
soloni (afvoices)				
Unfinetuned (v2)	44.47	36.07	29.61	20.24
Finetuned (v3)	**37.12**	**32.33**	21.17	**16.72**

Table 2: ASR Evaluation results: We apply the same normalization steps as explained in Section 3 and remove the tags from both the reference and the prediction before calculating the WER and CER. The values in bold highlight the best performance per metric for each benchmark.

Experimental setup: We used 4 NVIDIA A100 GPUs with 80GB VRAM each for these experiments. We finetuned `soloni-114m-tdt-ctc-v0` and `soloni-114m-tdt-ctc-v2` in this study. The two models have a hybrid architecture with a Fast-Conformer encoder (Rekesh et al., 2023) and two jointly trained decoders: an autoregressive TDT decoder, Token-and-Duration Transducer (Xu et al., 2023; Graves, 2012) and a convolutional decoder trained with a Connectionist Temporal Classification (CTC) loss function (Graves et al., 2006). We will refer to the models as soloni-v0 and soloni-v2. The finetuned versions of these models are identified on Hugging Face as v1 and v3, respectively.

Training Data: We finetuned the two models on the training set of the human-reviewed subset of Kunkado (33.47 hours). We simplified training for the ASR model by removing symbols and diacritics, and making number and Bambara-specific normalizations using our bambara-normalizer Python package. We then removed code-switching markers (the double underscores) and all punctuation. We kept only a reduced set of acoustic event tags (e.g., overlaps, paralinguistic vocalizations such as laughter, and music), which are modeled during training but ignored during evaluation. We applied these normalization steps after noticing that, with only just over 30 hours of training data, the models were struggling with the human-annotation variability present in the reference transcripts. These models rarely improved beyond 70% WER due to inconsistencies in numbers, tags and code-switching. We concluded that it would require much more data for an end-to-end ASR model to learn the original task. Table 3 contains 3 examples of how this normalization simplifies the task.

Training configurations: We trained soloni-v0 for 100k steps on 2 GPUs, with batch size 40, using the AdamW optimizer and Noam scheduler with a learning rate scaling factor of 0.03 and a 10% warmup ratio (Vaswani et al., 2017). We froze the encoder for the last 7,000 steps, training only the 5M combined parameters of the two decoders.

For soloni-v2, trained on much more data with the ANV dataset, we simply trained all the 114M parameters for 13k steps on 4 GPUs, with batch size of 64 and LR scaling factor of 1 and 3,000 warmup steps. Both models were trained with `bf16` float precision.

4 Evaluation & Results

We evaluated the Word Error Rate (WER) and Character Error Rate (CER) of the resulting models, and we also report the findings of the human evaluation conducted during the Bambara ANV project (Diarra et al., 2025a, Tall, 2025). Table 2 presents the WERs and CERs of the two models before and after our finetuning experiments; the terms in parentheses represent the dataset on which the unfinetuned versions were trained and their version IDs on Hugging Face. Since these models have two decoders, we only report the best scores; detailed per-decoder metrics can be found in their respective model cards. Both finetuned versions reduce WER on the Kunkado test set. Soloni-v3, our latest finetuned model, achieves the best results on both benchmarks[3].

In the detailed comparative analysis report by Tall on several RobotsMali ASR models, we found

[3]Note that, as one third of Nyana-Eval (15 examples) comes from the Kunkado test set, there is a small intersection between the two test sets

Original	Normalized
Bamananw ko ten ko maa jugu t'i ba sinamuso ye, nka n'a ni ba bɛ kɛlɛ la, <CHEVAUCHEMENT>__ voilà__ o kɔni ka di ye __ donc__ jamana...	bamananw ko ten ko maa jugu t'i ba sinamuso ye nka n'a ni ba bɛ kɛlɛ la <tag> voilà o kɔni ka di ye donc jamana
<MUSIQUE>an b'an sinsin ni Alahutala tɔgɔ barikama ye.	<MUSIQUE> an b'an sinsin ni Alahutala tɔgɔ barikama ye
ɛɛ... nɛgɛjuru sira 76 64 10 10... __76 64 10 10__	ɛɛ nɛgɛjuru sira bi wolonwula ni wɔɔrɔ bi wɔɔrɔ ni naani tan tan soixante-seize soixante-quatre dix dix

Table 3: Sample Kunkado transcripts pre and post normalization

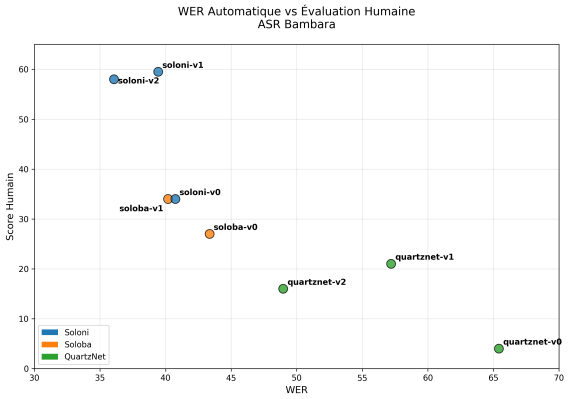

Figure 2: WER vs human evaluation. Figure from (Tall, 2025)

that soloni-v1 achieved the highest human evaluation score on Nyana-Eval (59.5 out of 135), outperforming soloni-v2, which was trained with an additional 98 hours from the African Next Voices Bambara dataset. Tall evaluated the model outputs on a 0–3 scale, where 0 indicated barely understandable transcriptions with multiple errors and 3 indicated semantically and lexically accurate transcriptions. Scores were assigned using criteria such as semantic and lexical fidelity, handling of code-switching and disfluencies, and accurate recognition of proper names. According to Tall, soloni-v1 shows clear gains in handling speech disfluencies, proper names, code-switching, and noisy or overlapping speech—phenomena that are common in natural Bambara conversations (Tall, 2025). The v3 models are not included in that report because they were finetuned from v2 after the report was published; however, the substantial WER improvement in soloni-v3 (32.33% vs. 36.07%) suggests that a proportional improvement in human scores is plausible. Figure 2 (from the comparative analysis report) ranks the models by both human scores and WER. All of these models, along with their v3 variants, are available on RobotsMali's Hugging Face profile.

5 Conclusion

In this work, we introduce Kunkado, a 160-hour Bambara ASR dataset compiled from Malian radio archives, "Media's Gift" to LRL Bambara NLP, and designed to better reflect present-day, naturally occurring speech. By shifting from curated source materials and controlled recording conditions toward broadcast content containing code-switching, disfluencies, background noise, and overlapping speech, we address a major source of domain mismatch that often yields models that perform well on curated data using standard metrics but poorly on real-world data and applications. Our experiments show that finetuning on 33.47 hours of human-reviewed Kunkado data yields substantial gains, and the best-performing configuration (soloni-v3) improves WER on both Kunkado-sourced test data and Nyana-Eval relative to its unfinetuned counterpart. These results support the broader conclusion that, for Bambara, representativeness and linguistic realism can matter as much as (or more than) raw hours when the goal is real-world usability. At the same time, the linguistic richness of spontaneous speech increases annotation and modeling difficulty, motivating pragmatic transcription guidelines and normalization choices, as well as continued investment in human review. We release Kunkado and the associated models to encourage research, community-driven quality standards led by native speakers, and future work on code-switching-aware evaluation and data collection and training from large-scale resources such as radio broadcasts for low-resource, predominantly-oral languages.

Limitations

The design of the Kunkado corpus intentionally prioritized the linguistic authenticity of the Bambara language. Specifically, the data reflects the most common and contemporary register of spoken Bambara, characterized by features inherent to natural conversational settings (e.g., podcasts, TV shows, and debates). These features include extensive code-switching, prevalent slang, and frequent spontaneous-speech disfluencies—exactly the features that the ground rules for the African Next Voices project required us to exclude from the dataset. (Diarra et al., 2025a)

Our ASR experiments demonstrated that these added linguistic complexities necessitate significantly greater volumes of data and advanced model engineering to achieve robust performance. Although radio broadcasts, in the West African region, represent a virtually inexhaustible, daily-generated source of data, resource constraints limited our labeling efforts to only ≈ 40 hours within the scope of this project, relying solely on limited internal funding.

The compelling results from the human evaluation indicate that continued annotation and development on more authentic data could substantially accelerate the deployment of high-fidelity, real-world speech applications for Bambara speakers. Moving forward, we advocate for a community-driven dataset design strategy where quality standards are organically defined with robust participation by native speakers.

Acknowledgments

We would like to extend our gratitude to Radio Benkouma, Mousso TV, ORTM and Radio Sahel FM for graciously sharing their archives with us and allowing us to release the data with an open source license.

References

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.

Brian DeRenzi, Anna Dixon, Mohamed Aymane Farhi, and Christian Resch. 2025. Synthetic voice data for automatic speech recognition in african languages. *Preprint*, arXiv:2507.17578.

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 Souleymane Coulibaly, Panga Azazia Kamaté, Madani Amadou Tall, Emmanuel Élisé Koné, Aymane Dembélé, and Michael Leventhal. 2025a. Dealing with the hard facts of low-resource african nlp. *Preprint*, arXiv:2511.18557.

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

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

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.

Christian Huber and Alexander Waibel. 2025. Handling numeric expressions in automatic speech recognition. *Preprint*, arXiv:2408.00004.

Yasmin Moslem. 2024. Leveraging synthetic audio data for end-to-end low-resource speech translation. *Preprint*, arXiv:2406.17363.

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.

James Robert. 2018. Pydub: Manipulate audio with a simple and easy high level interface.

Andrew Rosenberg, Yu Zhang, Bhuvana Ramabhadran, Ye Jia, Pedro J. Moreno, Yonghui Wu, and Zelin Wu. 2019. Speech recognition with augmented synthesized speech. *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 996–1002.

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

Chara Tsoukala, Kosmas Kritsis, Ioannis Douros, Nikolaos Kokkas, Vasileios Arampatzakis, Vasileios Sevetlidis, Stella Markantonatou, and George Pavlidis. 2023. Asr pipeline for low-resourced languages: A case study on pomak. pages 40–45.

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.

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

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.