Low-Resource Automatic Speech Recognition for Kinyarwanda

A. Mukamana, E. Habimana, J. Uwimana, C. Okoye · GCT Exchange, Kigali Hub · ACL Findings 2026

Abstract

We present **kinya-asr-base**, a 90M-parameter Conformer-CTC model trained on 1,200 hours of community-collected Kinyarwanda speech. The model achieves a word error rate of 12.4% on the public KinyaSpeech test split — a 38% relative improvement over the previous best — while remaining small enough to run on a $40 Android device. We release the model under Apache-2.0, the training corpus under CC-BY-SA 4.0, and a complete reproduction recipe. We discuss tone handling, code-switching with English, and three deployment lessons from a four-month pilot across two community clinics and one agricultural extension office in Eastern Province.

1. Introduction

Kinyarwanda is spoken by roughly 12 million people across Rwanda, eastern DRC, southern Uganda and parts of Tanzania. Despite this scale, public ASR systems perform poorly: commercial APIs we benchmarked in early 2025 ranged from 31% to 47% WER on conversational speech, and were effectively unusable on the elderly speakers most likely to need voice interfaces.

This paper documents a single attempt to close that gap in a way that is reproducible, auditable, and locally maintainable. Our contribution is not a new architecture — Conformer-CTC has been the workhorse of low-resource ASR for several years. The contribution is the data process, the tone handling, and the deployment discipline.

2. Data

KinyaSpeech-1.2k was collected over fourteen months in partnership with three community radio stations and the Kigali Public Library reading program. Speakers consented through a process co-designed with a community advisory panel. Recordings span read speech (412 hours), spontaneous interviews (538 hours), and call-in radio programs (250 hours). Speaker demographics are reported in Appendix A; we over-sample women and speakers over 55, who are systematically under-represented in existing African speech corpora.

3. Model

The acoustic model is a 12-layer Conformer encoder (d_model=512, 8 heads, kernel=31) with a CTC head over a 2,048-token unigram SentencePiece vocabulary. Total parameters: 90.4M. Training uses SpecAugment, speed perturbation (0.9, 1.0, 1.1), and a label-smoothed CTC loss. We train for 120k steps on 8 × A100 GPUs (roughly 41 GPU-hours).

4. Tone Handling

Kinyarwanda is a tonal language, but tone is not consistently written. In v0.3 we treated tone as out-of-band; the model collapsed on Eastern Province dialects where lexical tone disambiguates common

verb stems. In v0.4 we annotate a 40-hour subset with high/low tone markers derived from pitch tracking and use these as auxiliary CTC targets. WER on the Eastern Province subset improved from 19.1% to 13.6%.

5. Results

System	WER ↓	RTF ↓	Size
Commercial API A (2025)	31.2%	n/a	n/a
Commercial API B (2025)	47.0%	n/a	n/a
Prior open baseline	20.1%	0.41	240M
kinya-asr-base (ours)	12.4%	0.18	90M
kinya-asr-base INT8	13.1%	0.09	23M

6. Deployment

The INT8 variant runs at 0.09 real-time factor on a Tecno Spark 8C (Helio G35, 2GB RAM). We deployed an offline voice-note transcription tool at two community clinics in Bugesera and Nyagatare districts over four months. Clinicians used it for 31% of patient encounters; qualitative feedback emphasised the value of full offline operation given intermittent connectivity. A full deployment report is published separately.

7. Limitations & Ethics

The model performs poorly on children's speech (WER 24.8%), highly code-switched speech with English embedding (WER 18.2%), and recordings with significant background music. We do not recommend the model for forensic or evidentiary use. The training corpus excludes any recording flagged by community reviewers as sensitive, and a takedown process is documented in the dataset card.

8. Release

Model weights: huggingface.co/gctexchange/kinya-asr-base (Apache-2.0).
Dataset: huggingface.co/datasets/gctexchange/kinyaspeech-1.2k (CC-BY-SA 4.0).
Code: github.com/gctexchange/kinya-asr (Apache-2.0).
Reproduction recipe runs end-to-end on 8 × A100 in ~6 hours.

Acknowledgements

This work was supported by GCT Exchange research funds and an unrestricted grant from the Open Africa Compute Initiative. We thank the speakers, community reviewers, and the Kigali Public Library reading program.

References

[1] Gulati et al. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition.
[2] Park et al. 2019. SpecAugment.
[3] Kudo & Richardson 2018. SentencePiece.
[4] Nzeyimana 2023. KinyaBERT.
[5] African NLP community. 2024. Lessons from low-resource speech.