Multilingual Automatic Speech Recognition for Kinyarwanda, Swahili, and Luganda: Advancing ASR in Select East African Languages

Yonas Chanie, Moayad Elamin, Paul Ewuzie, Samuel Rutunda

Introduction

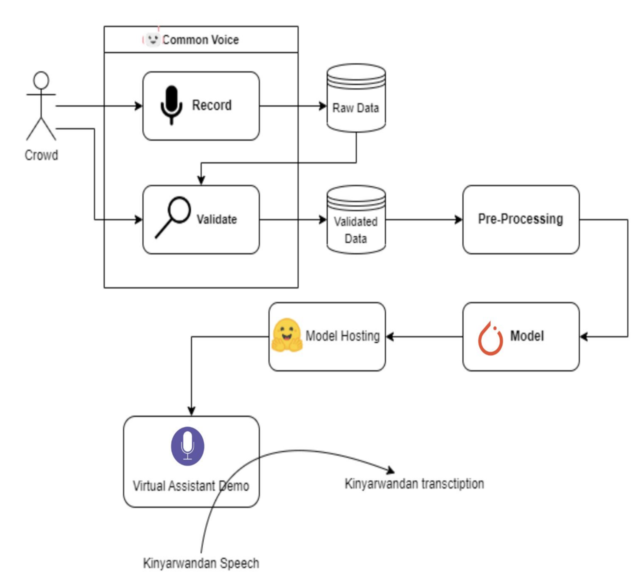
Automatic speech recognition (ASR) is a technology that enables computers to understand and transcribe human speech. This technology has a range of applications, such as voice-controlled devices including Alexa and Siri, transcription services, and accessibility tools for individuals with speech impairments.
Member of bantu language family, Kinyarwanda, Swahili and Luganda have over 90 million speakers across east Africa.

Objectives

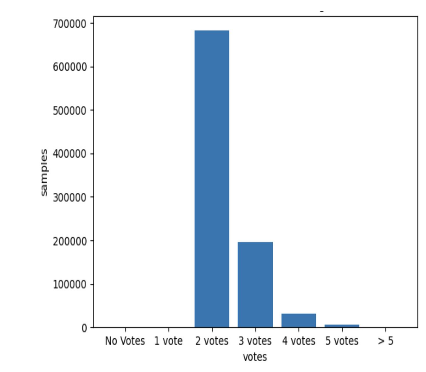
This project aims to address the following

- There is no robust automatic speech recognition system for Kinyarwanda, Swahili and Luganda
- Unavailability of quality data hinders the development of ASR based technologies
- Current ASR systems that support African languages are coupled with other major languages which causes performance degradation
- Unavailability of open-source models

Methodology

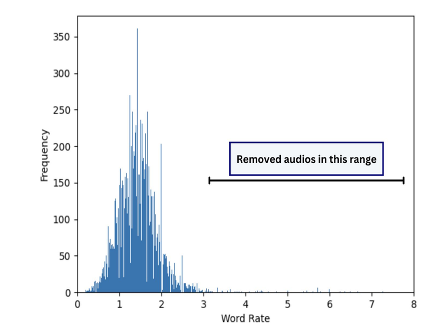


Word rate validation on Kinyarwanda



Analysis on the raw data[1] and the validated data showed equal number of samples with at least 2 upvotes

Raised concerns as regarding the veracity of the validation step



Filtered the validated data to retain only audios which has at most 3 words/sec

Text Preprocessing

Text Cleaning

- Lower Case
- Remove Special Characters (?:&_\\½√>€™$•¼})
- Normalize Apostrophes ('"′‘ -> ')
- Strip Accents from Loanwords (éèëêē, āâãá)
- Remove full stop at the end

Multilingual Data

To create a code-switched multilingual data, each sample from the languages were combined. A sample can have either multiple samples of a single language or randomized samples from the languages.

Data

Data split for the monolingual datasets (hours)

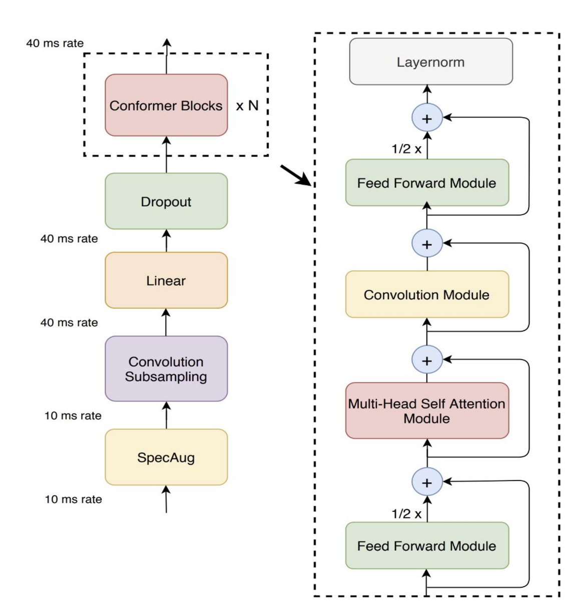
Split	Kinyarwanda	Swahili	Luganda
Train	1,284.8	48.59	102.1
Validation	24.3	16.7	20.1
Test	21.96	16.86	19.79

Data split for the code-switched multilingual dataset

Split	Full dataset (hours)	Size (samples)
Train	3,013.7	796,971
Validation	601.23	158,042
Test	300.64	79,053

Model Training

We trained on the *Conformer*[2] Speech to Text model using the cleaned dataset from our preprocessing steps



Tokenization
Byte-Pair encoding technique were used to tokenize the output text label.

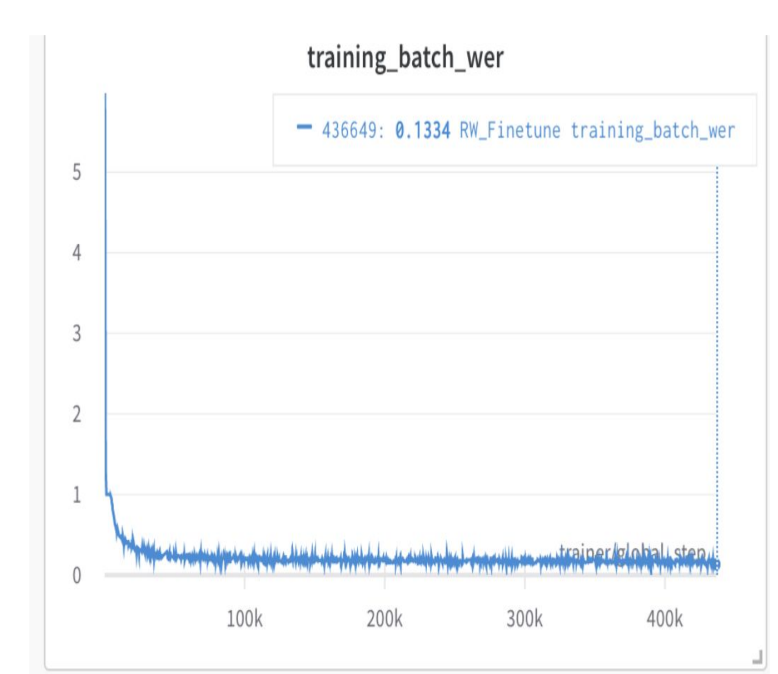
Evaluation
The results were evaluated using word error rate (WER)

$$\text{WER} = \frac{\text{Substitution} + \text{Insertion} + \text{Deletion}}{\text{Number of words in original text}}$$

Experiments

We trained on the following variations of the model for Kinyarwanda

- Train From Scratch
- Finetune Pretrained English Model
- Finetune Pretrained Kinyarwanda Model
- Train Medium Version



Results

We can see that the finetuned Kinyarwanda model performs better than the rest. The model is trained using the Mozilla Common Voice V9.0 dataset release and finetuned on the proposed dataset. This model has a WER of 13.34 on the training set after it is trained on 120 epochs.

We notice that there is a large gap between the value of the word error rate and character error rate. The property can be attributed to different factors including the quality of the audio, the complexity of the language, and the similarity in phonetics between different words that are written differently.

Models	Size (Parameters)	Model Size	Performance
From Scratch Model	121 Million	466MB	WER: 22.45 CER: 7.46
English Baseline Finetuned Model	121 Million	466MB	WER: 26.95 * CER: 8.64
Kinyawrwanda Baseline Finetuned Model	121 Million	466MB	WER: **17.57 (Test)** WER: **13.34 (Train)** CER: **5.28**
Medium Model	30 Million	119MB	WER: 24.20 CER: 7.99

Using the pre-trained Kinyarwanda conformer-based model, we finetuned the model on the proposed dataset to evaluate the performance on the test split of the dataset and baseline dataset splits.

Our result shows that while we were not able to improve the WER for Kinyarwanda compared to the monolingual baseline model, we can note that the CER is still on par and our multilingual model is able to correctly predict at the character level.

Test Set	WER	CER
Code-Switched	21.91	6.38
Kinyarwanda	25.48	7.79
Swahili	17.22	5.96
Luganda	21.95	5.15

Conclusion & Future Work

- **Data Inspection and processing** were instrumental in achieving impressive results (WER 5.0 with Original model)
- **Engaging a Linguist** for further data validation will be helpful to identify missed language errors in data
- **Finetune** with recent models like Whisper[3]
- **Hyperparameter Tuning:** *Better Accuracy*
- **Model Compression:** *Low-end devices*

References

1. https://commonvoice.mozilla.org/en/datasets
2. Gulati, Anmol, et al. "Conformer: Convolution-augmented transformer for speech recognition." *arXiv preprint arXiv:2005.08100* (2020).
3. Radford, Alec, et al. "Robust speech recognition via large-scale weak supervision." *arXiv preprint arXiv:2212.04356* (2022).

Demo

Contact:
ychanie@andrew.cmu.edu
melamin@andrew.cmu.edu
pewuzie@andrew.cmu.edu
samuel@digitalumuganda.com

giz Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH
FAIR FORWARD
DFKI German Research Center for Artificial Intelligence
Digital Umuganda
MASAKHANE

Carnegie Mellon University Africa