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

![Graph showing word rate validation on Kinyarwanda](image)

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)

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.

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.

<table>
<thead>
<tr>
<th>Models</th>
<th>Size (Parameters)</th>
<th>Model Size</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>From Scratch Model</td>
<td>121 Million</td>
<td>460MB</td>
<td>WER: 22.45 * CER: 7.46</td>
</tr>
<tr>
<td>English Baseline Finetuned Model</td>
<td>121 Million</td>
<td>460MB</td>
<td>WER: 26.95 * CER: 8.64</td>
</tr>
<tr>
<td>Kinyarwanda Baseline Finetuned Model</td>
<td>121 Million</td>
<td>460MB</td>
<td>WER: 21.91 * CER: 7.99</td>
</tr>
<tr>
<td>Medium Model</td>
<td>30 Million</td>
<td>119MB</td>
<td>WER: 24.20 * CER: 7.99</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fast Set</th>
<th>WER</th>
<th>CER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Code-Switched</td>
<td>25.48</td>
<td>7.79</td>
</tr>
<tr>
<td>Kinyarwanda</td>
<td>21.95</td>
<td>5.79</td>
</tr>
<tr>
<td>Swahili</td>
<td>17.22</td>
<td>5.96</td>
</tr>
<tr>
<td>Luganda</td>
<td>21.95</td>
<td>5.15</td>
</tr>
</tbody>
</table>

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


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

Data split for the monolingual datasets (hours)

<table>
<thead>
<tr>
<th>Split</th>
<th>Train</th>
<th>Validation</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio</td>
<td>1:284.8</td>
<td>102.1</td>
<td>20.1</td>
</tr>
<tr>
<td>Text</td>
<td>21.95</td>
<td>18.68</td>
<td>9.19</td>
</tr>
</tbody>
</table>

Data split for the code-switched multilingual dataset

<table>
<thead>
<tr>
<th>Split</th>
<th>Full dataset (hours)</th>
<th>Size (GB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>3,033.7</td>
<td>798.971</td>
</tr>
<tr>
<td>Validation</td>
<td>601.23</td>
<td>158.042</td>
</tr>
<tr>
<td>Test</td>
<td>300.64</td>
<td>79.053</td>
</tr>
</tbody>
</table>

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

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

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

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