Abstract

- **BembaSpeech**: 24 hours of read speech
- An exploration of approaches for end-to-end Bemba ASR
- Multilingual 1 billion XLS-R model gives the best result of 32.91% WER.

Motivation

- Need of a speech dataset for the Bemba language to build ASR systems.

Bemba Language

- **Bemba** (also referred to as ChiBemba, Icibemba) is a Bantu language principally spoken in Zambia by over 30% of the population.
- It is a written but low resourced language.
- It has 5 vowels and 19 consonants
- Writing system: Latin script.

Description of BembaSpeech

- 24 hrs read speech.
- 14, 438 utterances.
- 17 speakers; 9 male & 8 females.
- Fixed splits: train, dev & test sets with no speaker overlap
- Audio files encoded in *.wav format.
- Sample rate: 16kHz.

Characteristics of the BembaSpeech ASR

<table>
<thead>
<tr>
<th>Subset</th>
<th>Size (Hrs)</th>
<th># of Files</th>
<th>Speakers</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>20</td>
<td>11906</td>
<td>8</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Dev</td>
<td>2.5</td>
<td>1556</td>
<td>7</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Test</td>
<td>2</td>
<td>977</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>24.5</td>
<td>14438</td>
<td>17</td>
<td>9</td>
<td>8</td>
</tr>
</tbody>
</table>

Table 1: General characteristics of the BembaSpeech ASR dataset

Data Collection Methodology

- Audio recording tool: Lig-Aikuma Mobile App[1].
- Text sources: diverse publicly available sources/domains including; Bemba literature books and youtube video transcripts.
- Preprocessed the dataset to ensure data accuracy.
- Availability: publicly available under CC BY-NC-ND 4.0 license.

Experiments and Results

- E2E model: DeepSpeech (v0.8.2)
- Dataset size: 17 hrs subset of dataset.
- LM1: Transcripts (train and dev sets)
- LM2: Transcripts + JW300
- Baseline model - training from scratch using DeepSpeech.
- Monolingual pre-trained model: DeepSpeech English model
- Multilingual pre-trained model: Wav2vec2.0 based XLS-R [300 million and 1 billion parameter models]
- The 1 billion XLS-R model achieves the best result of 32.91% WER

Future work

- Carry out in-depth error analysis
- Try the Transformer based language model
- Investigate unusual results
- Expand the size of the corpus

References