ParlaSpeech-HR – a Freely Available ASR Dataset for Croatian Bootstrapped from the ParlaMint Corpus

Nikola Ljubešić, Danijel Koržinek, Peter Rupnik, Ivo-Pavao Jazbec

ParlaCLARIN 3 workshop
Marseille, 20 June 2022
Overview

- Background data
- Dataset construction process
- Dataset description
- ASR experiments
- Future work
Background data

- Manual transcripts
  - ParlaMint 1 corpus (version 2.1)
  - Croatian is represented with data from one term (2016-2020), 20 million words

- Video / audio data
  - Harvested from the parliamentary YouTube channel
  - 2,419 hours of speech data
Dataset construction process (1)

- No open ASR model, very bad acoustic model
- Google speech-to-text has a reasonable ASR model for Croatian (word-error-rate 27.4%), and a $300 voucher
- Let us
  a. Transcribe some data with Google STT and align the automatic transcription with the manual transcription (Plüss et al. 2019)
  b. Learn an in-house model on the data from (a) and transcribe and align all available data
Dataset construction process (2)

**long speech - text alignment**

1. Audio → speech recognition
   - acoustic models
   - language model
   - Adapt at each iteration

2. text-text alignment
   - transcription

3. text-aligned speech

Repeat for unaligned audio and text segments

https://sail.usc.edu/old/software/SailAlign/
Dataset description

- Dataset obtained with Google STT – 66 hours
- Final dataset – 1,816 hours, 403,925 segments of 8-20 seconds in length (perfect for ASR training)
- Metadata on 309 speakers from ParlaMint – name, gender, age, party, party status
- train:dev:test split, in test three separate male and three female speakers
- [http://hdl.handle.net/11356/1494](http://hdl.handle.net/11356/1494) (CC-BY-SA 4.0)
## ASR experiments

<table>
<thead>
<tr>
<th>System</th>
<th>WER</th>
<th>CER</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM/WFST baseline</td>
<td>66.92%</td>
<td>50.43%</td>
</tr>
<tr>
<td>GMM/WFST adapted</td>
<td>30.54%</td>
<td>12.60%</td>
</tr>
<tr>
<td>TDNN/WFST</td>
<td>22.51%</td>
<td>9.78%</td>
</tr>
<tr>
<td>TDNN/WFST chain</td>
<td>16.38%</td>
<td>6.91%</td>
</tr>
<tr>
<td>XLS-R-66-initial</td>
<td>13.94%</td>
<td>5.42%</td>
</tr>
<tr>
<td>XLS-R-110-original</td>
<td>10.57%</td>
<td>3.23%</td>
</tr>
<tr>
<td>XLS-R-110-normalized</td>
<td>10.15%</td>
<td>3.04%</td>
</tr>
<tr>
<td>XLS-R-300</td>
<td>7.61%</td>
<td>2.34%</td>
</tr>
<tr>
<td>Slavic-300</td>
<td>6.79%</td>
<td>2.22%</td>
</tr>
<tr>
<td>Slavic-300+lm</td>
<td>4.30%</td>
<td>1.88%</td>
</tr>
</tbody>
</table>
Future work

- Speaker profiling benchmark - identity, gender, age, political orientation
- Datasets in other languages – Czech and Polish inside ParlaMint 2, but other, less resourced languages are very welcome to join the ParlaSpeech family!
- Multimodal corpora available through concordancers (video, gesture recognition?)