Montreal Forced Aligner: an accurate and trainable aligner using Kaldi
Michael McAuliffe¹, Michaella Socolof², Sarah Mihuc², Michael Wagner¹, & Morgan Sonderegger¹

¹Department of Linguistics, McGill University ²Department of Linguistics, University of Maryland

Most speech systems are designed to be used for a specific language and are not easily adapatible for other languages. This project aims to provide an accurate and trainable aligner using Kaldi.

**BACKGROUND**

**Featured Tools**
- Kaldi
- HTK
- FAVE-align
- MAUS
- EasyAlign
- Gentle

**Features**
- Kaldi-based
- Trainable
- Tested on 20+ languages
- Can model words not in the dictionary
- Preserves alignments of other words
- Triphone acoustic models
- Right and left context for phones (models coarticulation)
- Acoustic features adapted by speaker
- More accurate alignment
- Parallel processing helps scaling up
- Command line interface
- Well-tested, easy-to-use
- Actively maintained
- Well-documented and open source
- Input
  - Orthographic TextGrid and label files
  - Wav files
- Output
  - Aligned TextGrids

**REFERENCES**

[Goldman 2011]

**EVALUATION**

**How do alignments from the Montreal Forced Aligner compare with a state-of-the-art system?**

- Read speech from production experiment (48 minutes)
  - “Please say ___ again”
  - Target 1-2 syllable words with vowel + obstruent
  - Vowel and obstruent of target word were hand annotated
  - Vowel begin, vowel end, and obstruent end
  - Force aligned
- Compared with Prosodylab-aligner
  - Also trainable
  - Uses similar acoustic models to other systems (Monophone GMM)
- Conditions:
  - Flat – Trained on limited data (48 minutes)
  - Pretrained on lab recordings (15 hours)
  - Pretrained on LibriSpeech (474 hours)

**INPUT**

**SYSTEM COMPARISON**

<table>
<thead>
<tr>
<th>System</th>
<th>Toolkit</th>
<th>Trainable</th>
<th>Acoustic model</th>
<th>Pretrained models</th>
<th>Supported platforms</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFA</td>
<td>Kaldi</td>
<td>Yes</td>
<td>Triphone GMM</td>
<td>English</td>
<td>Mac, Linux, Windows</td>
</tr>
<tr>
<td>Prosodylab-aligner²</td>
<td>HTK</td>
<td>Yes</td>
<td>Monophone GMM</td>
<td>English, French</td>
<td>Mac, Linux</td>
</tr>
<tr>
<td>FAVE-align/P2FA³</td>
<td>HTK</td>
<td>No</td>
<td>Monophone GMM</td>
<td>English</td>
<td>Mac, Web, Windows</td>
</tr>
<tr>
<td>(Web)</td>
<td>HTK</td>
<td>Non-trivial</td>
<td>Monophone GMM</td>
<td>English + 8 other languages</td>
<td>Linux, Web</td>
</tr>
<tr>
<td>MAUS⁴</td>
<td>HTK</td>
<td>No</td>
<td>Monophone GMM</td>
<td>English + 3 other languages</td>
<td>Windows</td>
</tr>
<tr>
<td>EasyAlign¹</td>
<td>HTK</td>
<td>No</td>
<td>Monophone GMM</td>
<td>English</td>
<td>Mac, Web</td>
</tr>
<tr>
<td>Gentle³</td>
<td>Kaldi</td>
<td>No</td>
<td>ANN</td>
<td>English</td>
<td>Mac, Web</td>
</tr>
</tbody>
</table>

**OUTPUT**

**RESULTS**

**DISCUSSION**

- Montreal Forced Aligner outperforms the Prosodylab-Aligner
- Pretrained models on larger datasets are generally preferable than only using the dataset to be aligned
- Larger data sets may be unnecessary if the style/reading conditions are the same