Forced Alignment of Spoken Audio

Josef Fruehwald
19 April 2016
Why Forced Alignment?
What we had

Data - Static
What we wanted:

Data - Dynamic

dob


i:
Getting from what we have to what we want

1. Convert analogue recordings to digital format.
   - Preserve the most important metadata
2. Identify where in the audio speech sounds of interest are.
3. Automate the acoustic analysis of the speech sounds.
4. Apply statistical analysis to the acoustic analysis for inferences.
Identifying where in the audio speech sounds interest are.

"Forced Alignment"
Finding words in audio

A bar chart showing the frequency of stems across different days of the week.
Forced Alignment

Rest of the presentation:

- What some of the necessary bits and pieces are for doing forced alignment.
- What some of the tools out there are for doing alignment as an end user.
Bits and Pieces and Issues for doing forced alignment
## Piece 1: A pronouncing dictionary

<table>
<thead>
<tr>
<th>word</th>
<th>pronunciation</th>
</tr>
</thead>
<tbody>
<tr>
<td>well</td>
<td>W EH1 L</td>
</tr>
<tr>
<td>there</td>
<td>DH EH1 R</td>
</tr>
<tr>
<td>was</td>
<td>W AH0 Z</td>
</tr>
<tr>
<td>one</td>
<td>W AH1 N</td>
</tr>
<tr>
<td>time</td>
<td>T AY1 M</td>
</tr>
</tbody>
</table>
Issue 1: Pronunciation

Variants

What do you do for multiple pronunciations? e.g. Bailey (2016)

<table>
<thead>
<tr>
<th>word</th>
<th>pronunciation</th>
</tr>
</thead>
<tbody>
<tr>
<td>walking</td>
<td>W AO1 L K IH0 N</td>
</tr>
<tr>
<td>walking</td>
<td>W AO1 L K IH0 NG</td>
</tr>
<tr>
<td>walking</td>
<td>W AO1 L K IH0 NG G</td>
</tr>
</tbody>
</table>
Issue 1: Pronunciation

Variants

Option 1: Include all options
Let the aligner figure out which option to use.

- **Pros**
  - You'll get more accurate timing.

- **Cons**
  - In choosing pronunciation variants, some aligners have a lower rate of agreement with humans coders than humans coders do with each other (Bailey 2015)
  
  - It can be tricky to identify which pronunciations are variants of each other.
Issue 1: Pronunciation

Variants

Option 2: Only include one option
Only allow the aligner to choose one option

· Pros
  - It'll be easier to identify all instances of potential pronunciation variation.

· Cons
  - The timing information will be less accurate.
Issue 2: Out of Dictionary Words

No matter how large a pronouncing dictionary you're working with, there will always be some words in free flowing speech that aren't in the dictionary.

<table>
<thead>
<tr>
<th>word</th>
<th>pronunciation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fruehwald</td>
<td>F R UW1 W AO0 L D</td>
</tr>
<tr>
<td>hoagie</td>
<td>HH OW1 G IY0</td>
</tr>
</tbody>
</table>

These either need to be added to the dictionary when the aligner is run, or a separate piece of software needs to try to guess the pronunciation based on the spelling.
Piece 2: An acoustic model

\[ [w] = \]
Piece 3: A transcript

Outside of the original fieldwork, this is the most time consuming and expensive part.
How it works

How it works

beginning middle end beginning middle end
How it works

if \( p > p \), beginning
if \( p < p \), middle
How it works

\[ \ddot{o} \quad \ddot{a} \]

beginning  beginning  middle  end  beginning  middle  middle  end
Concerns about forced alignment

It'll make mistakes

- It is easier and faster (read: cheaper) to manually correct the output of automated systems than to create the annotations from scratch.

- Humans make mistakes too! And the kinds of mistakes automated systems make are usually *systematic*, so they're easier to identify and locate.
Concerns about forced alignment

It's a black box!
You are a black box
Concerns about forced alignment

Automation removes me from the data
Doing Forced Alignment at Home
FAVE

The FAVE-suite is actually two pieces of software: An aligner, and a Bayesian formant analyzer.

- Aligner based on p2fa, trained on 25 hours of US Supreme Court oral arguments.
- Fairly good time accuracy.

![Histogram of forced alignment errors.](image)

Fig. 3. Histogram of forced alignment errors.
FAVE Benefits

- Developed assuming that multiple talkers in the audio was the default case.
- Developed in the open, trying to be as cross-platform friendly as possible.
- Written in Python, which is a very widely understood programming language.
- The system is relatively simple and flexible (although its acoustic models are not).
- The primary developer is friendly and responsive 😊
FAVE Cons

- Based on North American acoustic models, although MacKenzie & Turton have found it compares favorably to other aligners on British data.

<table>
<thead>
<tr>
<th></th>
<th>Median</th>
<th>Mean</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Onset</td>
<td>Offset</td>
<td>Onset</td>
</tr>
<tr>
<td>FAVE</td>
<td>0.009</td>
<td>0.009</td>
<td>0.019</td>
</tr>
<tr>
<td>PLA</td>
<td>0.015</td>
<td>0.019</td>
<td>0.267</td>
</tr>
<tr>
<td>SPPAS</td>
<td>0.150</td>
<td>0.155</td>
<td>0.504</td>
</tr>
</tbody>
</table>
Recommended FAVE Usage

Download and install locally

Extensive documentation online, written assuming minimal familiarity with command line interfaces.

Welcome to the FAVE wiki!

Both FAVE-align and FAVE-extract are a collection of Python scripts. As such, you’ll need to download and install Python before using them. The FAVE suite was developed in Python 2.x, so we recommend installing the most recent Python 2.x version (2.7.6 as of writing).

FAVE-align

- Installing FAVE-align
- Using FAVE-align
What FAVE needs as input

- Audio
- Transcriptions
  - Partially time aligned
  - Multiple speakers annotated separately
Prosodylab Aligner

Developed at University of McGill, Montreal

Pros & Cons

- Much the same as FAVE, but re-training of the acoustic models is built in.
- No streamlined facility yet for multiple talkers
Prosodylab Aligner

Recommended Usage

- Download & Install
webMAUS

Developed in association with CLARIN-D

· Web-based platform
  - Easy to use
  - Less easy to adapt to task specific purposes
  - May be tricky if there are ethics restrictions on where and how your data is stored.

· No multiple talkers yet
Recommended Usage
webMAUS

Recommended Usage

![webMAUS Interface]

The image shows the webMAUS interface with the Recognizer section highlighted. The Recognizer is set to BAS – WebMAUS Basic. The Parameters section includes a setting for Language of the input.
DARLA

System developed at Dartmouth University

• Pros
  - Includes an automatic speech recognition system.

• Cons
  - So far, just a web-based service, with servers in the US
Recommended Usage

DARLA: Dartmouth Linguistic Automation
A suite of vowel formant extraction tools tailored to research questions in sociophonetics.

Completely Automated Vowel Extraction
Audio → Formants
This automated method is designed for "big data" research projects where a general vowel space is prioritized over word-level transcription accuracy.
The End