Forced Alignment of Spoken Audio

Josef Fruehwald

19 April 2016

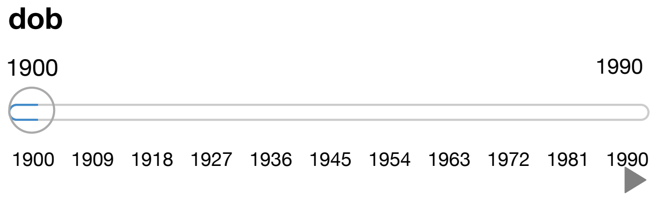
Why Forced Alignment?

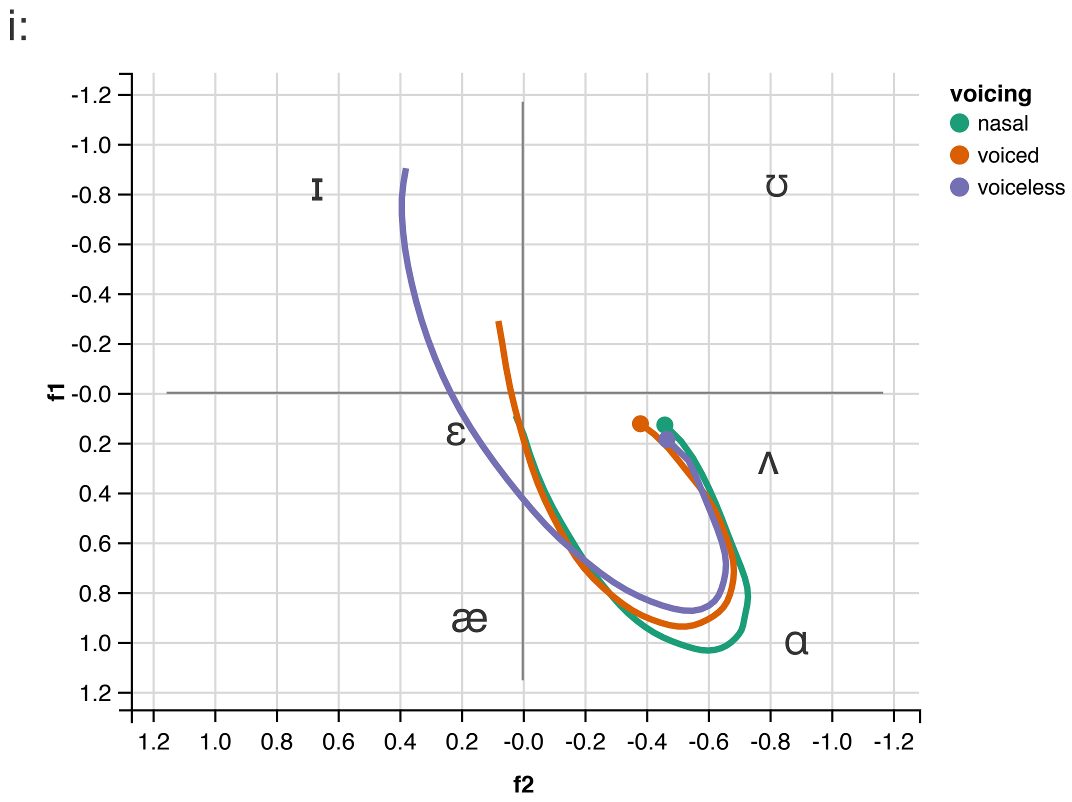
What we had

Data - Static

What we wanted:

Data - Dynamic



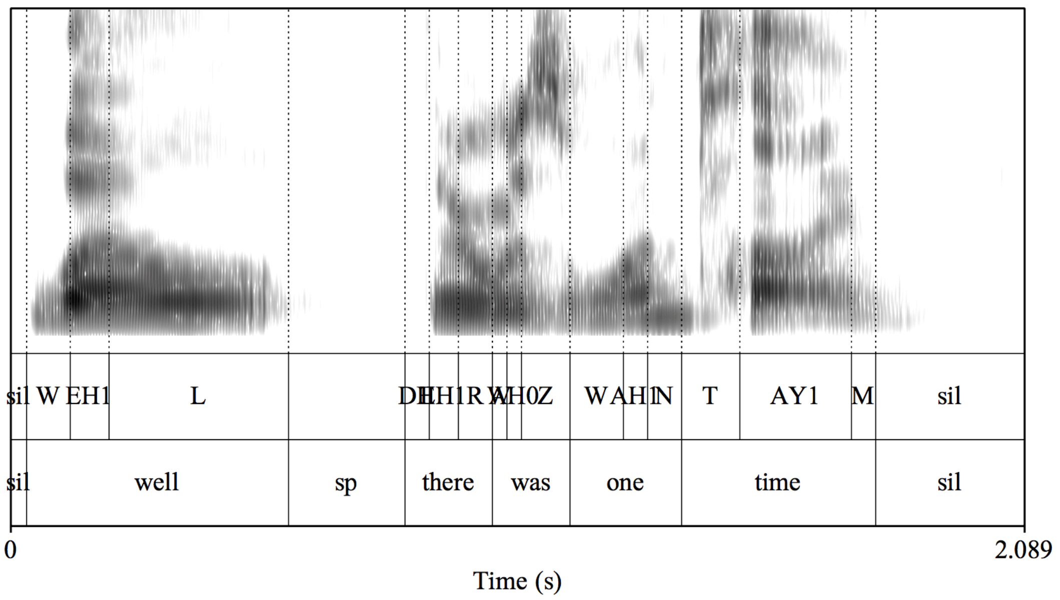


Getting from what we have to what we want

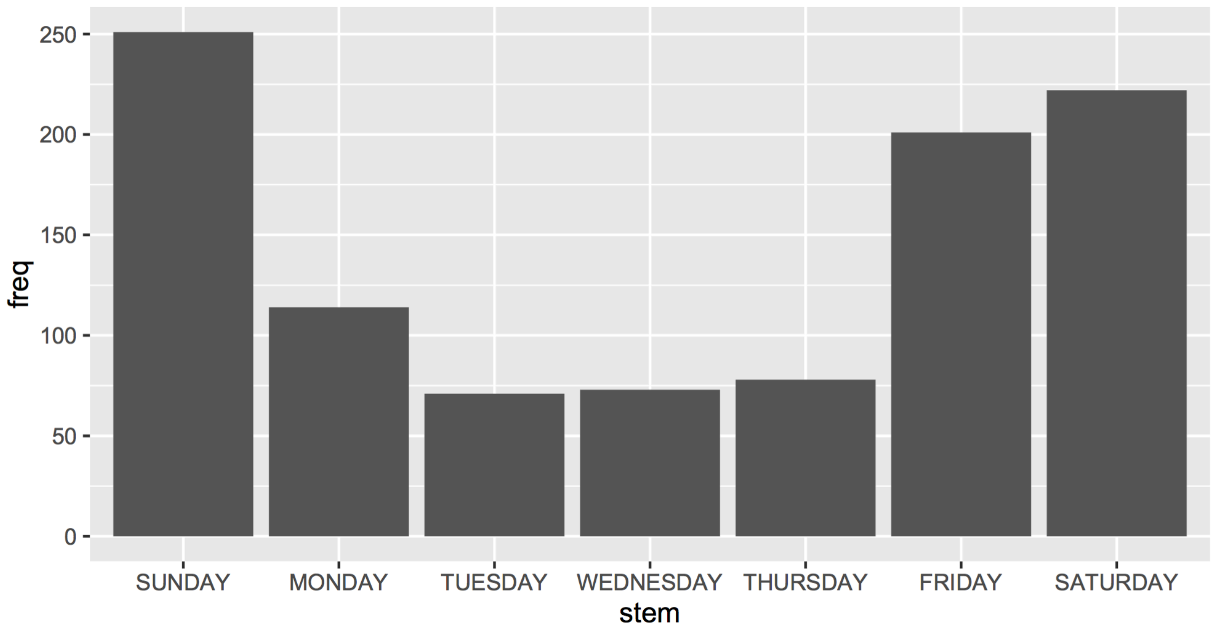
1. Convert analogue recordings to digital format.
    - Preserve the most important metadata
2. **Identify where in the audio speech sounds of interest are.**
3. Automate the acoustic analysis of the speech sounds.
4. Apply statistical analysis to the acoustic analysis for inferences.

Identifying where in the audio speech sounds interest are.

"Forced Alignment""



Finding words in audio



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

word	pronunciation
well	`W EH1 L`
there	`DH EH1 R`
was	`W AH0 Z`
one	`W AH1 N`
time	`T AY1 M`

Issue 1: Pronunciation Variants

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

word	pronunciation
walking	`W AO1 L K IH0 N`
walking	`W AO1 L K IH0 NG`
walking	`W AO1 L K IH0 NG G`

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

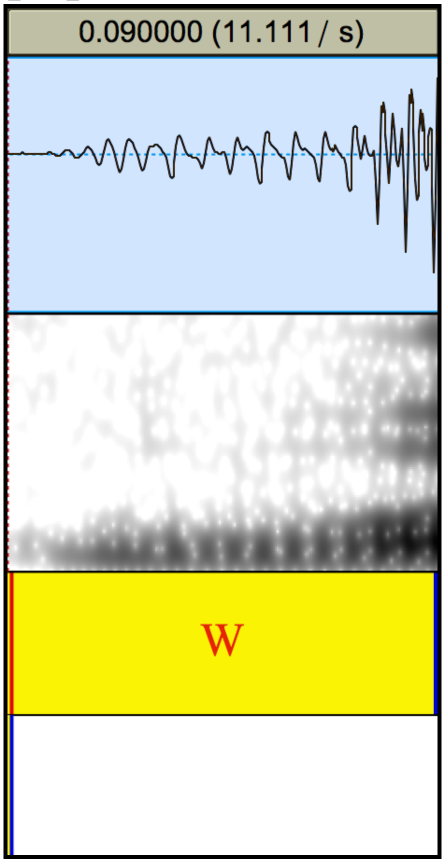
No matter how large a pronouncing dictionary you're working with, there will always be some words in free flowing speech that aren't in the dictionary.

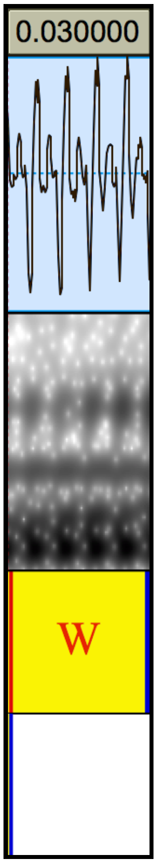
word	pronunciation
Fruehwald	`F R UW1 W AO0 L D`
hoagie	`HH OW1 G IY0`

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

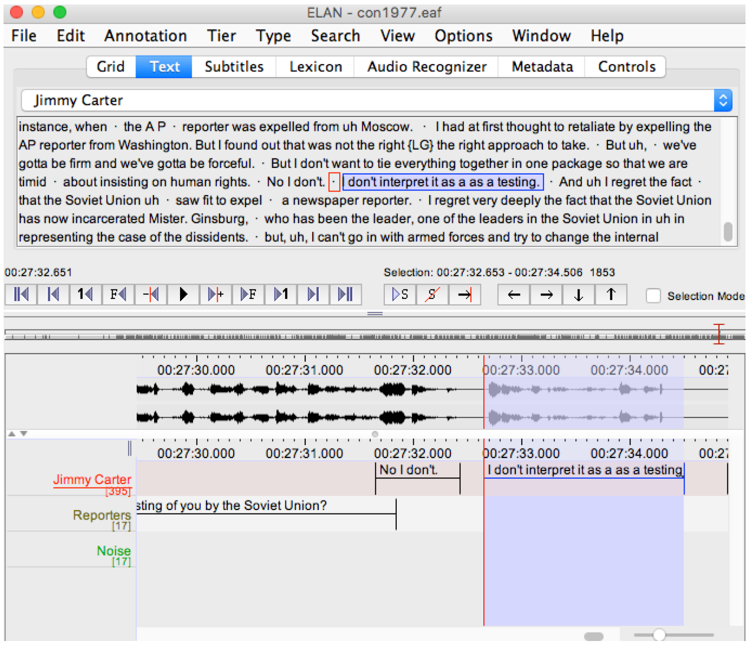
[w]=





Piece 3: A transcript

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


ELAN - con1977.eaf
File Edit Annotation Tier Type Search View Options Window Help
Grid Text Subtitles Lexicon Audio Recognizer Metadata Controls
Jimmy Carter
Selection Mode
Jimmy Carter [395]
Reporters [17]
Noise [17]

instance, when · the A P · reporter was expelled from uh Moscow. · I had at first thought to retaliate by expelling the AP reporter from Washington. But I found out that was not the right {LG} the right approach to take. · But uh, · we've gotta be firm and we've gotta be forceful. · But I don't want to tie everything together in one package so that we are timid · about insisting on human rights. · No I don't. · I don't interpret it as a as a testing. · And uh I regret the fact · that the Soviet Union uh · saw fit to expel · a newspaper reporter. · I regret very deeply the fact that the Soviet Union has now incarcerated Mister. Ginsburg, · who has been the leader, one of the leaders in the Soviet Union in uh in representing the case of the dissidents. · but, uh, I can't go in with armed forces and try to change the internal

00:27:32.651

Selection: 00:27:32.653 - 00:27:34.506 1853

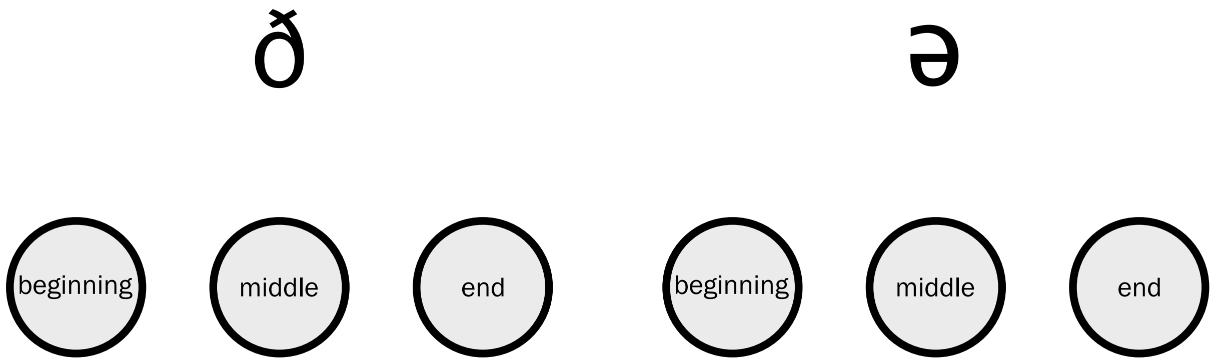
00:27:30.000 00:27:31.000 00:27:32.000 00:27:33.000 00:27:34.000

No I don't.

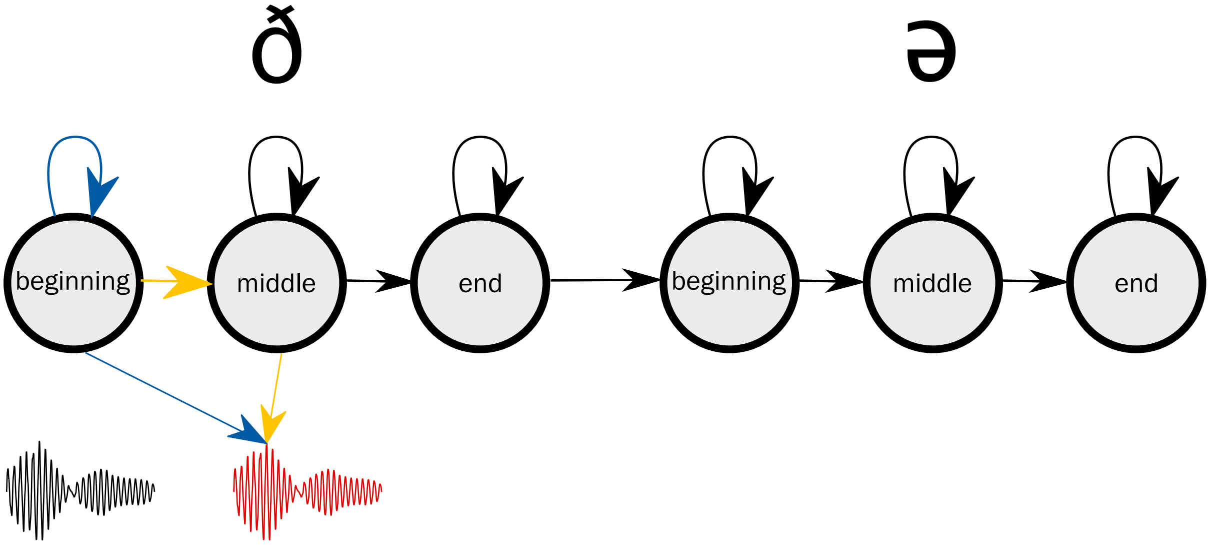
I don't interpret it as a as a testing

sting of you by the Soviet Union?

How it works

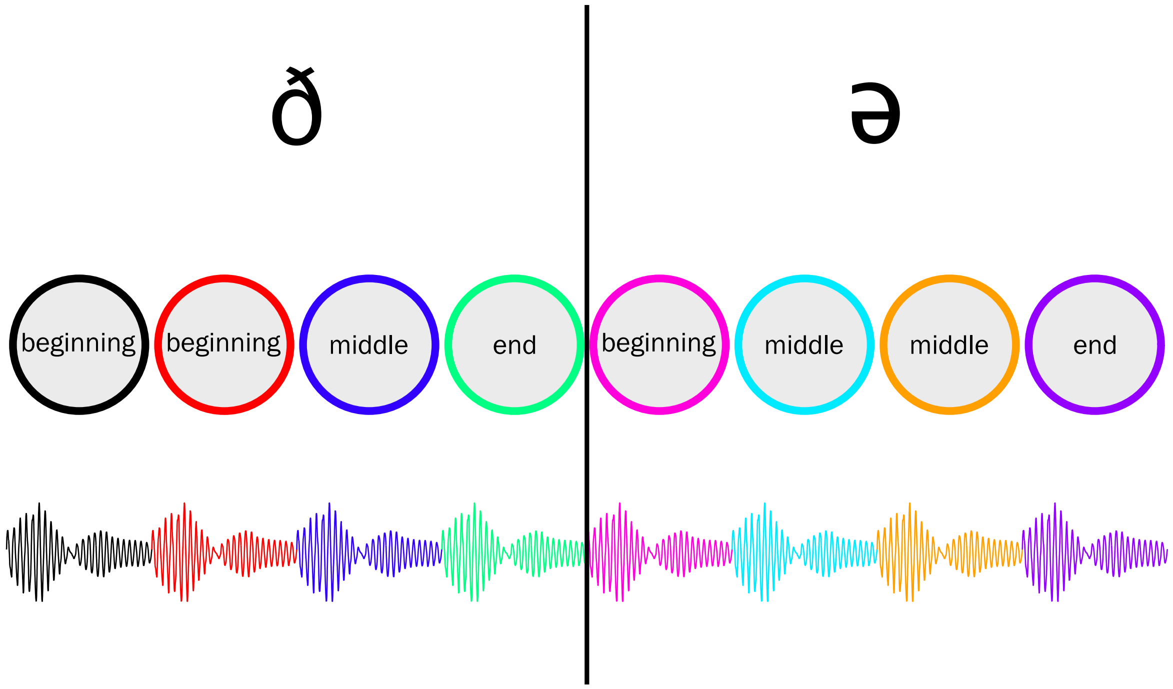


How it works



if **p** > p, beginning
if **p** < p, middle

How it works



ð
ə
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 sytems make are usually *systematic*, so they're easier to identify and locate.

Concerns about forced alignment

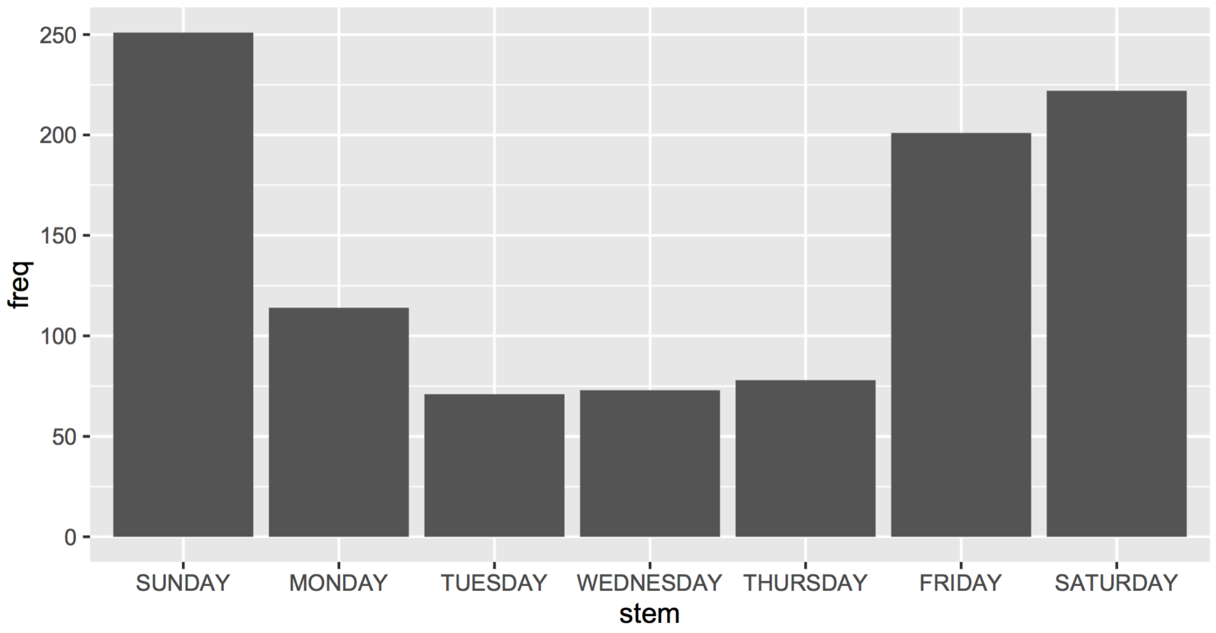
It's a black box!

You are a black box

Phrenological Chart of the Faculties.

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.

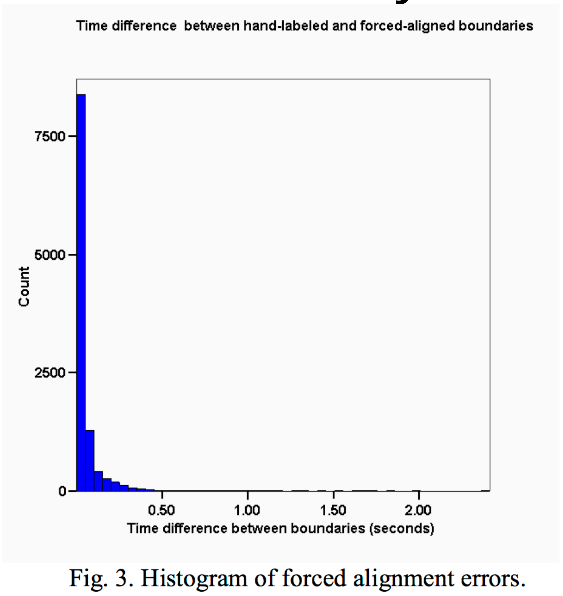


Fig. 3. Histogram of forced alignment errors.

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

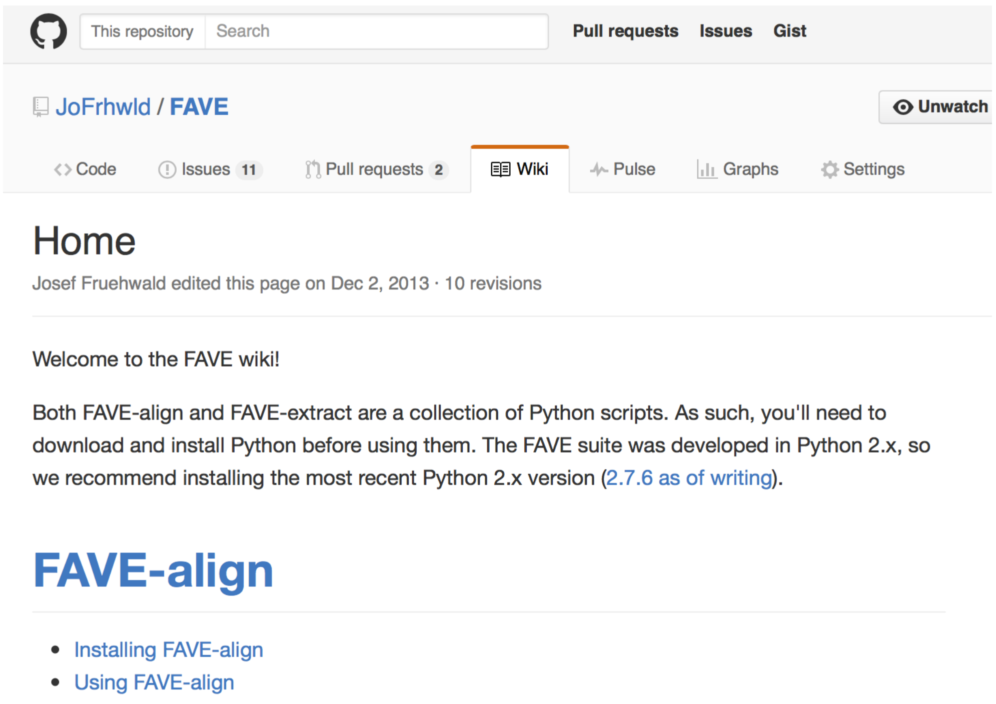
- Based on North American acoustic models, although MacKenzie & Turton have found it compares favorably to other aligners on British data.

	Median		Mean		Max	
	Onset	**Offset**	**Onset**	**Offset**	**Onset**	**Offset**
FAVE	0.009	0.009	0.019	0.021	0.583	0.588
PLA	0.015	0.019	0.267	0.252	55.473	55.488
SPPAS	0.150	0.155	0.504	0.480	68.903	67.408

Recommended FAVE Usage

Download and install locally

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



JoFrhwld / FAVE

Home

Josef Fruehwald edited this page on Dec 2, 2013 · 10 revisions

Welcome to the FAVE wiki!

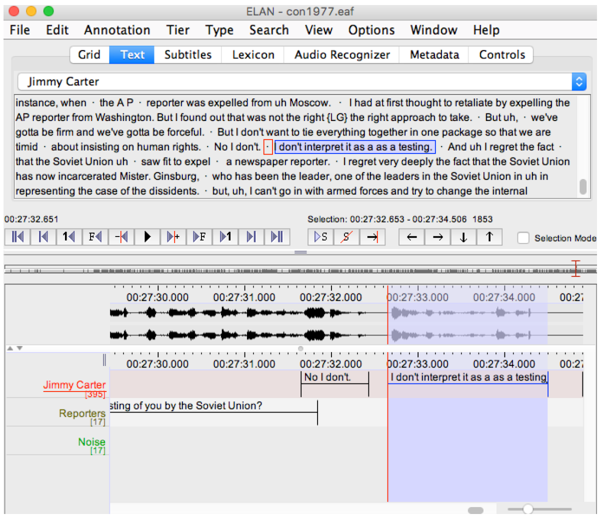
Both FAVE-align and FAVE-extract are a collection of Python scripts. As such, you'll need to download and install Python before using them. The FAVE suite was developed in Python 2.x, so we recommend installing the most recent Python 2.x version (2.7.6 as of writing).

FAVE-align

- Installing FAVE-align
- Using FAVE-align

What FAVE needs as input

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


ELAN - con1977.eaf
File Edit Annotation Tier Type Search View Options Window Help
Grid Text Subtitles Lexicon Audio Recognizer Metadata Controls
Jimmy Carter

instance, when · the A P · reporter was expelled from uh Moscow. · I had at first thought to retaliate by expelling the AP reporter from Washington. But I found out that was not the right {LG} the right approach to take. · But uh, · we've gotta be firm and we've gotta be forceful. · But I don't want to tie everything together in one package so that we are timid · about insisting on human rights. · No I don't. · I don't interpret it as a as a testing. · And uh I regret the fact · that the Soviet Union uh · saw fit to expel · a newspaper reporter. · I regret very deeply the fact that the Soviet Union has now incarcerated Mister. Ginsburg, · who has been the leader, one of the leaders in the Soviet Union in uh in representing the case of the dissidents. · but, uh, I can't go in with armed forces and try to change the internal

00:27:32.651

Selection: 00:27:32.653 - 00:27:34.506 1853

No I don't.

I don't interpret it as a as a testing

sting of you by the Soviet Union?

Prosodylab Aligner

Developed at University of McGill, Montreal

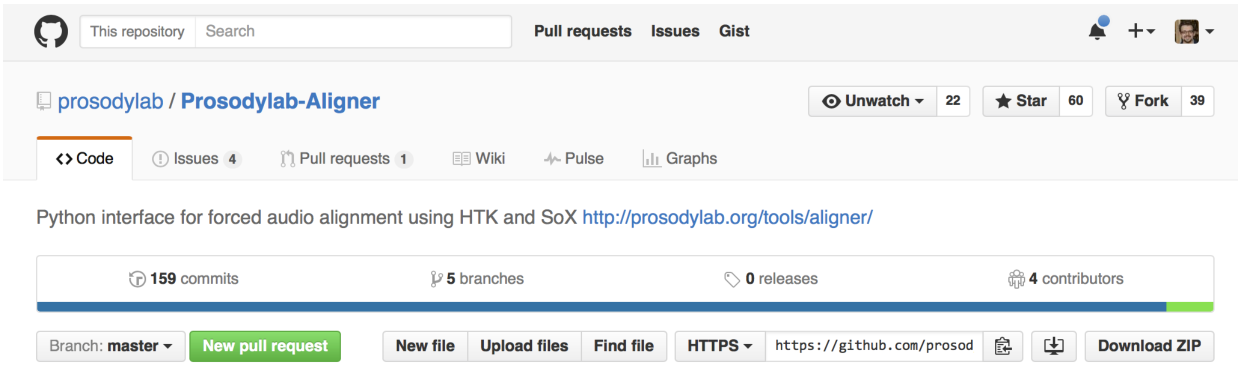
Pros & Cons

- 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



This repository | Search | Pull requests | Issues | Gist

prosodylab / **Prosodylab-Aligner**

Unwatch 22 | Star 60 | Fork 39

Code | Issues 4 | Pull requests 1 | Wiki | Pulse | Graphs

Python interface for forced audio alignment using HTK and SoX http://prosodylab.org/tools/aligner/

159 commits | **5** branches | **0** releases | **4** contributors

Branch: **master** | New pull request | New file | Upload files | Find file | HTTPS | `https://github.com/prosod` | Download ZIP

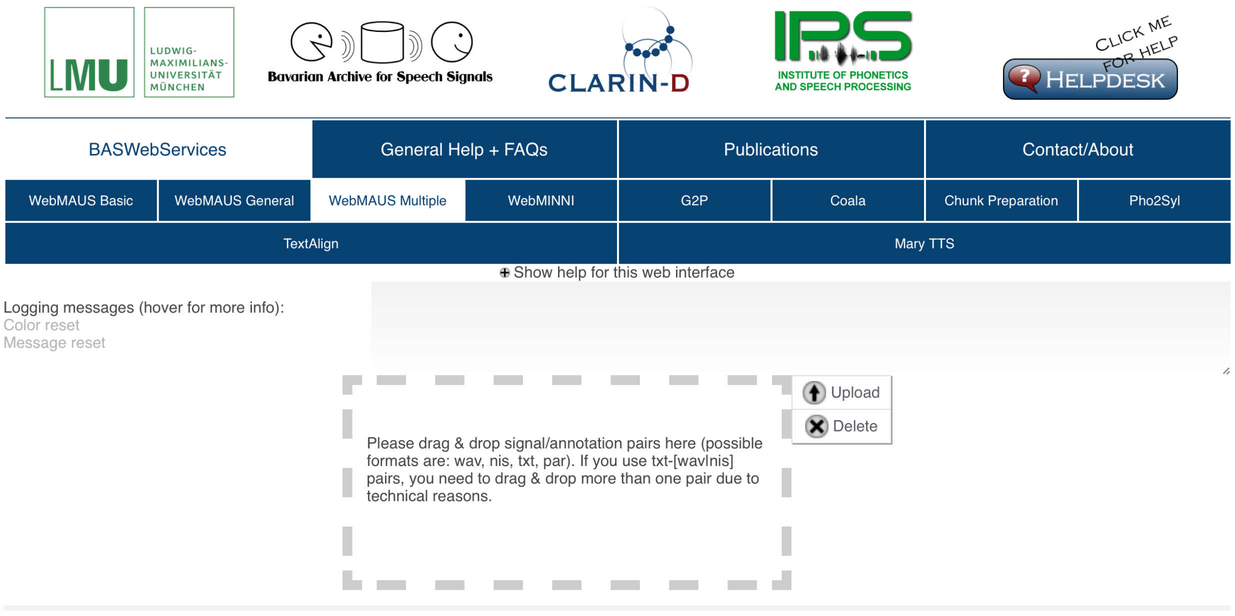
webMAUS

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

webMAUS

Recommended Usage

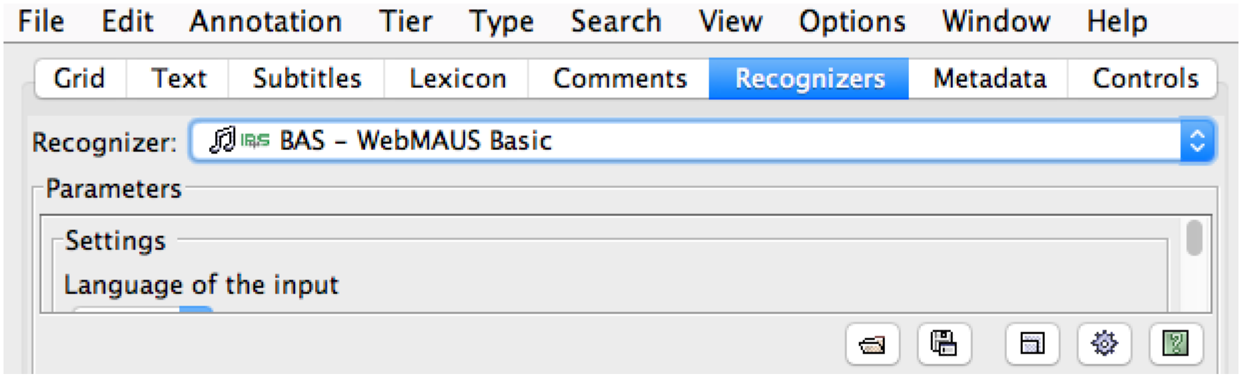


Logging messages (hover for more info):
Color reset
Message reset

Please drag & drop signal/annotation pairs here (possible formats are: wav, nis, txt, par). If you use txt-[wav|nis] pairs, you need to drag & drop more than one pair due to technical reasons.

webMAUS

Recommended Usage


File Edit Annotation Tier Type Search View Options Window Help
Grid Text Subtitles Lexicon Comments Recognizers Metadata Controls

Recognizer: BAS – WebMAUS Basic

Parameters

Settings

Language of the input

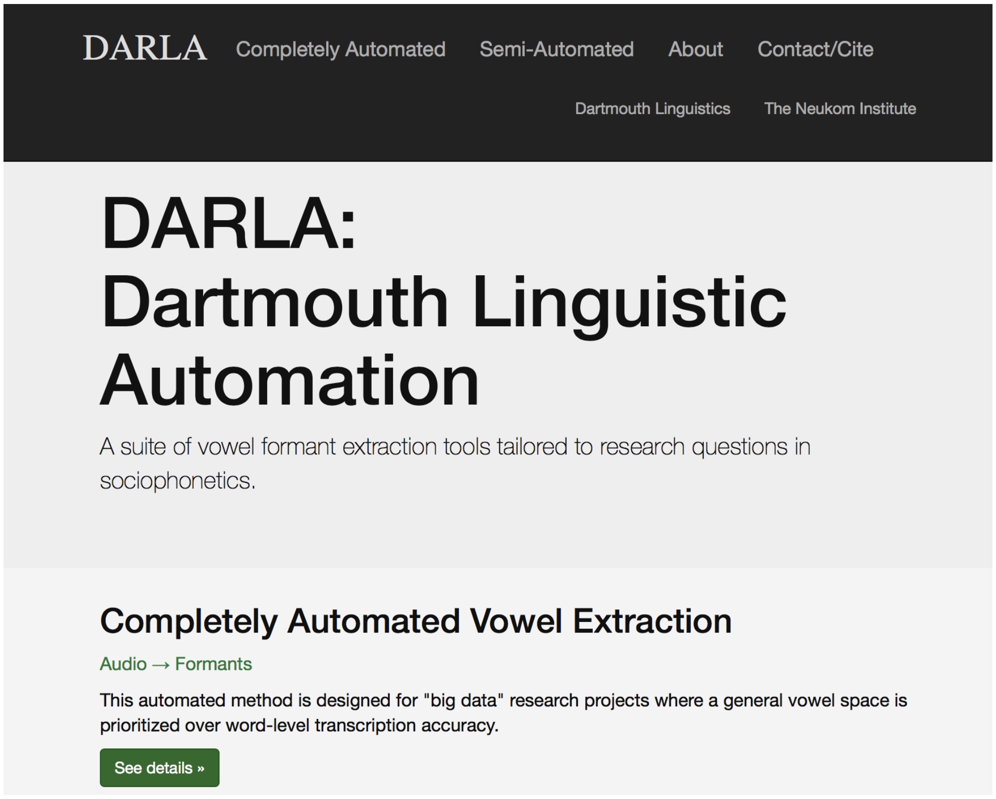
DARLA

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

DARLA

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