Forced Alignment in Linguistic Research: Workshop and Applications

Thursday, May 12th, 2016 - University of Texas, Austin

Eric Wilbanks - North Carolina State University
http://ericwilbanks.github.io/workshops.html

Today's Workshop

- Overview of Forced Alignment
- Using FAVE for English Data
- - Break -
- Using FASE for Spanish Data
- Misc. Advanced Topics and Resources

Today's goal is to give you the tools to use force-alignment in your own research. We'll be working with some data that I've prepared as well as your own data throughout.

But First,

Who Am I?

- Sociophonetician working on English and Spanish variation (ask me after the workshop!)
- My goal is to give you the tools you need to do linguistic research, not to convert you into programmers

Who Are You?

- Interested in processing spoken data and learning new methodologies
- No prior CS or coding experience presupposed
- I make take some general concepts from acoustic phonetics as shared knowledge, but just stop me whenever you have a question or want a clarification. (I love questions!)

Getting Started

We'll be using the following resources today, so please make sure you have them installed or downloaded.

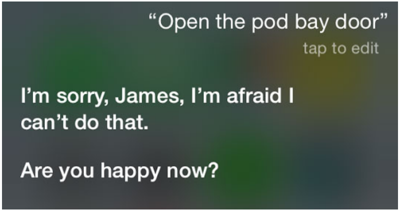
- Praat - `http://www.fon.hum.uva.nl/praat/`
- Example wav files and transcripts
  `http://ericwilbanks.github.io/workshops.html`

Excellent!

Let's get started!

Speech Recognition

- Speech Recognition technologies are becoming ubiquitous in our world.
- Fully automated phone-level transcription is still a bit far off. (Though see DARLA: `http://darla.dartmouth.edu/`)
- But linguists can still benefit from these technologies!



“Open the pod bay door”

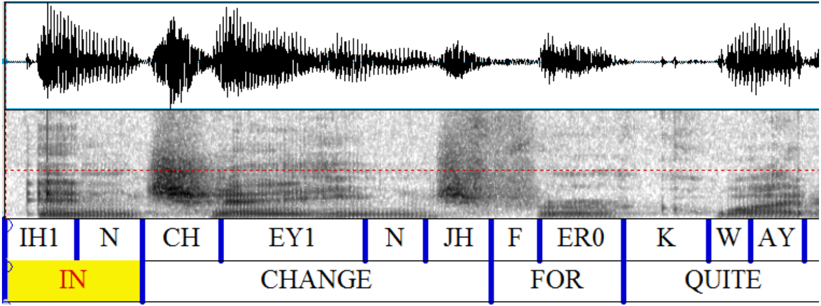
I’m sorry, James, I’m afraid I can’t do that.

Are you happy now?

Speech Recognition

- Goal of Speech Recognition: determine sequences of words given a sound input
    1. What words tend to cooccur? - Language Model
    2. What do parts of words (usually phonemes) “look” like?
- Forced alignment takes smaller, easier problem:
    - We know the order of the words, where are the boundaries between phones?

End Goal



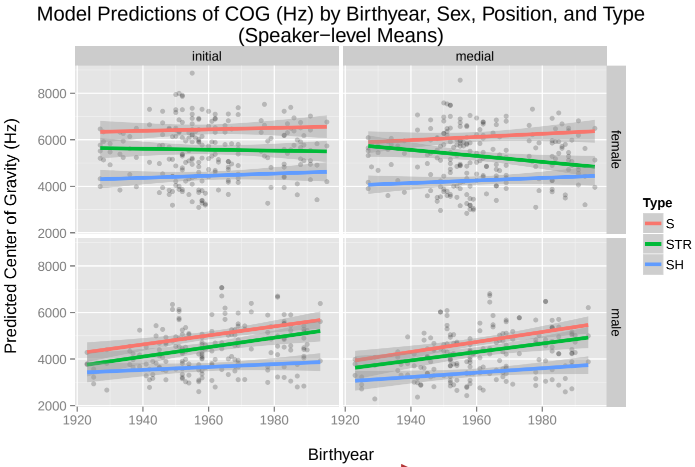
Automatic Phone-Level Transcriptions

- Manual segmentation of phones is incredibly time-consuming, at some estimates 800x real-time (Schiel and Draxler, 2003).
- Automated segmentation, however, is increasing by orders of magnitude the amount of acoustic data linguists are able to analyze.
- As Labov et al. (2013) note, utilizing forced alignment allowed them to increase tokens extracted from each interview from 300 to 9,000.

Raleigh Example - (str) Retraction

Female b. 1961	Female b. 1991
other part of the street	live down the street
street	street
s	s

Raleigh Example - (str) Retraction



- 140 force-aligned speakers from Raleigh
- 99,150 tokens used in modeling.

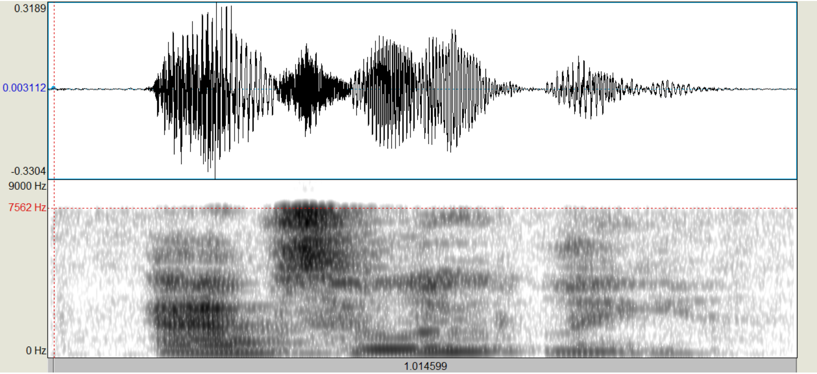
Typical Work Flow

1. Collect your Data
2. Transcribe
3. Force-Align
4. Extract measurements via Praat scripts
5. Analyze in R, Stata, Excel, etc.

Under the Hood

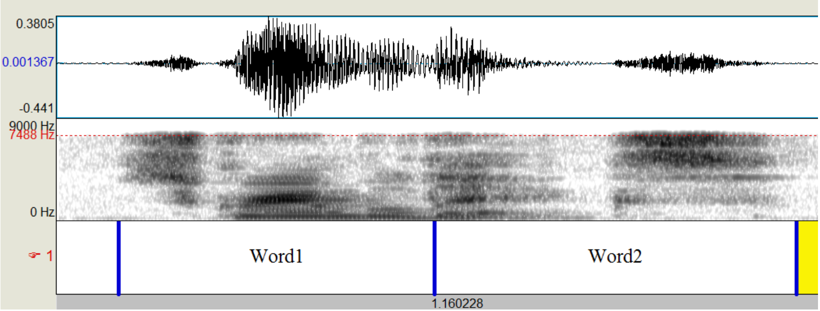
- We're going to talk now a little bit about how Forced Alignment systems actually work.
- Don't worry, no math knowledge required!
- It's important to have a general knowledge of what's happening so we can better interpret the output or deal with possible errors.

Let’s Pretend We’re Siri



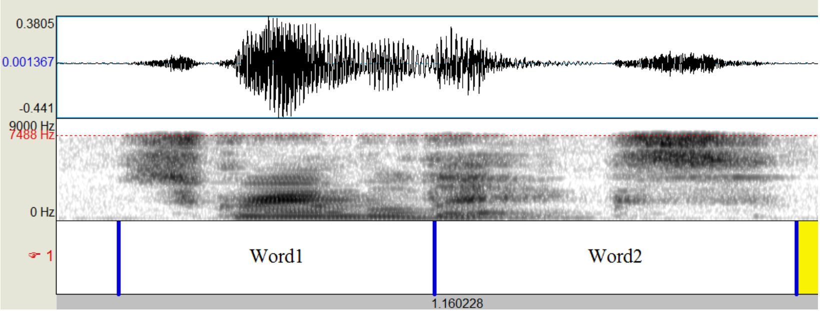
What word is this? ((Mystery Word))

Probabilistic Knowledge



This is tougher. What if I told you..

Probabilistic Knowledge



This is tougher. What if I told you..

Word1 is "Scrambled"? ((Mystery Phrase))

Recognition vs. Alignment

- That was really hard. It's hard for computers too.
- Luckily, we don't have to guess about the words when doing forced-alignment.
- All the computer has to figure out is where the boundaries go between segments.

What does a forced-alignment system need to function?

Dictionary

Given a word, what phonemes should I be looking for?

- Dictionary creation can be incredibly time-consuming.
- Luckily, we have the hand-made CMU Pronouncing Dictionary for English (Weide, 1994).
- Depending on the language in question, you might be able to go from orthography to phonemic representation automatically.
- *Potato - Potato?* – Multiple entries

Phoneme Models

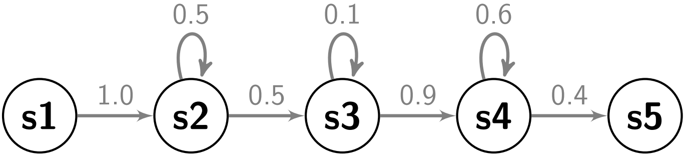
How do I teach a computer what /t/ looks like?

- Lots of variability in speakers, allophones, environments etc.
- Need to capture time-dynamics: an /s/ can be 30ms or 2000ms.
- Has to be able to be computationally-feasible.

Hidden Markov Models

- Hidden Markov Models (HMMs) take a sequence of observations (in our case acoustic vectors) and give them some label (phones)
- This is done by modeling each label/phone as a sequence of "hidden" states.
- During training, observations are paired with labels so that transition probabilities between states and model vectors can be learned.

Left-To-Right HMM Model of Phone



These 5 states represent a single phone, let's say /s/.

During training, statistics are gathered on the acoustic information present at each state.

We train the model to learn the characteristics of each state and the probability of being in a given state given what the spectrum is doing at that point.

Possible Challenges

Considering what we now know about how forced-aligners function underneath the hood, namely:

1. Finite dictionary mapping words to sequences of phones
2. Linear models of phonemes based on acoustic observations

What difficulties could a system like this have?

Bad Transcriptions

- It's very easy to filter out repeats or false starts during transcription.
- If a word is in the transcript but not in the sound file, alignment of neighboring words can suffer.
- Similar problems occur when a word isn't in the transcription but is in the sound file.
- Luckily, these sorts of problems are easy to spot and fix. Just adjust the transcript and re-align.

Unknown Words

- CMU dictionary is extensive, but there are words it doesn't include.
- You'll have to add missing entries like this to your custom dictionary by hand.
- Currently, the custom dictionary we use for the Raleigh project is around 3.6k entries.
- These days, each new interview usually only adds 3-6 new words.

Overlap and Noise

- While humans are really good at comprehending overlapping voices, computers are not yet.
- Having overlapping voices, or a voice and some noise in the same frequency range can often confuse the aligner.
- After all, we taught it what /s/ looks like, not what a simultaneous /s/ and /æ/ looks like
- FAVE has a nice way of dealing with this, we'll talk about it later.

Code-Switching/Mixing

- Multilingual speech recognition systems are still a ways away.
    - Consider English “Pot o’ tea” [paɾəti] vs. Spanish “Para ti” [paɾati]
- Forced-alignment systems are currently only set up to align one language at a time.
- If you only have a little bit of code-switching, easiest is usually just to “hack” it by adding a custom dictionary entry for the code-switch.
- If there’s a lot of code-switching, you might have to split up your interview by language and then use two separate systems (if they both exist!)

Challenging Segments

- If you've done hand-segmentation, you know that some segments are just impossible to segment.
- Consider the case of contiguous sibilants: "this shop" or worse "this scene"
- Vowel/Liquid Combinations are the absolute worst.
- These segments often suffer from poor inter-rater reliability when humans segment.
- Benefit of using forced-alignment: consistency of boundaries. Given the same acoustics, the boundary will always be placed in the same position.

Review!

Forced Aligners Work By

1. Matching words in transcript to entries in the dictionary to create sequence of phones
2. Using HMMs of phones and the acoustics to determine where boundaries should be placed.

Because of this, some difficulties include:

- Bad transcriptions
- Unknown words
- Overlap and noise
- Code-Switching
- Generally tough segments

Using FAVE

Forced Alignment and Vowel Extraction

FAVE - Background

- The first large-scale forced-aligner in linguistics was P2FA (Penn Phonetics Lab Forced Aligner) (Yuan and Liberman, 2008).
- These English acoustic models are built on recordings of US Supreme Court Justices
- Models are quite robust, most widely used models for English
- FAVE website adds automatic vowel extraction options as well.

FAVE - Options

- **Website**
  - Convenient and easy interface
  - Can be slow, especially if there's a lot of traffic.
  - Should Penn experience a server outage, you'll be out of luck.
  - Need to ensure that your IRB covers the type of data transmission here.
- **On Your Computer**
  - More convenient, quicker, and you don't have to rely on Penn's servers being up, etc.
  - Requires you to install HTK `http://htk.eng.cam.ac.uk/`
  - This is tough and requires a bit of technical know-how.

Web Interface

- Today we'll be using files I've already prepared to learn about the interface and aligning process.
- Then, we'll practice with some of the English data you have.
- Please download the files listed at `http://EricWilbanks.github.io/workshops.html`
- All data presented in this section are from the Buckeye Corpus (Pitt et al., 2007) or CABank (CallFriend) (Yaeger-Dror et al., 2004)
- Then, open up FAVE-align's website `http://fave.ling.upenn.edu/FAAValign.html`

Basic Example

- Let's get our hands dirty with some alingning!
- Try aligning the `example.txt` and `example.wav` files
- This is a straight-forward clip that is high quality with no overlap or missing words.

Out of Dictionary Words

- Now, try aligning with the `example_missing.txt` and `example_missing.wav` files.
- Don't change any of the options in the interface and look at the output.
- Notice that the aligner skipped a word it didn't have in its dictionary ('Scooney')
- Now, run the aligner with the "-u Check transcription for unknown words" option.
- This will give you a list of the words the dictionary doesn't know. This is usually a good first step for any file.
- With the list of unknown words in hand...

Custom Dictionary Entries

- We'll have to add this unknown word ('Scooney') to our custom dictionary.
- Open up a notepad or equivalent and save it as a .txt file
- Dictionary entries are in the following format:
- SCOONEY `(tab)` S K UW1 N IY2
- Where `(tab)` represents an actual tab
- ARPABET symbols used for phone mapping can be found here: `https://en.wikipedia.org/wiki/Arpabet`

Out of Dictionary Words

- Now, go back to FAVE-align and reselect the files.
- This time, check "-i Import dictionary transcriptions" and select your custom dictionary file
- Compare this alignment to the previous one.

Overlap

- Open up the example_overlap TextGrid and WAV files in Praat and listen to them.
- Note the extensive overlap!
- This time, align using the example_overlap.txt and example_overlap.wav files.
- Since FAVE has speaker turn begin/end, it can align overlapping parts separately.

It’s time to work with your own data!

Generating Tab File

- Recall that FAVE requires a specific 5-tab format for transcriptions
- To generate these from praat textgrids, use the praat script FAVE provides (`http://fave.ling.upenn.edu/downloads/Convert_To_FAVE-align_Input.praat`)

Generating Tab File

To generate these tab files from SRT style files:

- Use the following python script by **Rachel Tatman**.
- This will create a Praat TextGrid.
- Use FAVE's praat script to create the 5-column tab-file.

It's a bit of a workaround at the moment. Might be developing more efficient options.

Generating Tab File

To generate these from CLAN transcriptions, do the following:

1. Open CLAN, set working directory, and in the Command box type:
   - `FLO +d1 myfile.cha +fS`
2. This will create a txt file of the format required by FASE

- The newer functionality of `FLO` is only included in CLAN versions after 1/26/16, so you may have to update
- `http://childes.psy.cmu.edu/clan/`

Using FASE

Forced Alignment System for Español

FASE Background

- FASE is a Spanish Forced-Aligner I've been working on off and on for a year and a half. (Thanks Jeff Mielke, Jim Michnowicz, and Rebecca Ronquest for assistance & data!)
- It's trained on speakers mainly from Mexico (sorry, no thetheo!) from the Corpus de Español de Raleigh-Durham (CERD) at NC State (`https://sites.google.com/a/ncsu.edu/michnowicz/research/cerd`).
- Labor of love == Not all features I'm planning are implemented yet.

Monophone Inventory

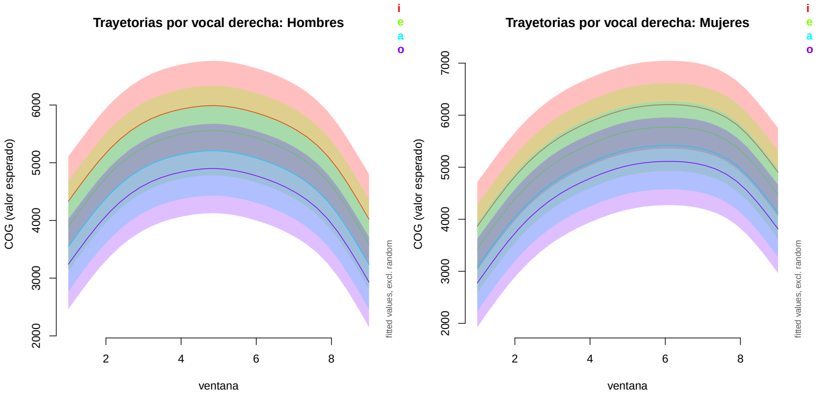
	labial	dental	alveolar	palatal	velar
plosives - voiceless	p	t			k
plosives - voiced	b	d			g
fricatives - voiceless	f		s		x
fricative - voiced				j (y)	
affricate				tʃ (CH)	
nasals	m		n	ɲ (NY)	
lateral			l		
rhotic - tap			ɾ (r)		
rhotic - trill			r (R)		

Vowels: /a,e,i,o,u/ correspond to their ipa symbols
Non-Speech: Laughing {LG}, Coughing {CG}, Breath {BR}, Noise {NS}, short pause (sp), silence {SIL}

Dictionary Construction

- The dictionary was constructed from the 44 million words **SUBTLEX-ESP** corpus (Cuetos et al., 2011).
- English loan words removed from corpus by cross-referencing with CMU Pronouncing Dictionary (Weide, 1994) and manually sorting.
- Final Spanish Pronunciation Dictionary - 93,350 unique words

Application Example



2,805 intervocalic /s/ from 20 speakers

/i/ → /e/, /a/, /o/

Web Interface

- Let's look at the web interface now
- Data from this section comes from the Spanish in Texas (SpinTX) corpus `https://spanishintexas.org/`

`http://phon.chass.ncsu.edu/cgi-bin/webalign_fase.cgi`

FASE Example

- Open up the `spanish_example` files in Praat and a text editor.
- Let's align it and look at the output
- Keep in mind the strategies we learned in the previous section!

Features to Come

- Addition of training data from SpinTX
- Freely definable speaker labels
- Stress marking in dictionary
- Overlap strategy FAVE has implemented
- Downloadable toolkit
- (Further Future) - Bilingual functionality, switching between FAVE and FASE models; automatic mapping of orthographic to phonemic
- Stay tuned, FASE may be relocating to a new website in the future!

It’s time to work with your own data!

Installing HTK on your Computer

- If you plan on using force-aligning fairly frequently, it's a really good idea to have an installation of HTK on your computer.
- Both FAVE and FASE (as well as almost all other forced-alignment systems out currently) require HTK
- HTK is open-source for non-commercial purposes (hooray!) but can be a pain to install; luckily it's a one time thing.
- Instructions for installation can be found at:
    - `http://htk.eng.cam.ac.uk/docs/inst-nix.shtml`
    - `https://github.com/JoFrhwld/FAVE/wiki/Installing-FAVE-align`

Other Aligners

- **Prosodylab Aligner** (Gorman et al., 2011) provides models for NA English and Quebec French and also supports training of novel models.
- **SPLaligner** (Milne, 2014) French aligner trained on Canadian political recordings
- **PraatAlign** (Lubbers and Torreira, 2015) Praat plugin with support for a variety of languages
- **EasyAlign** (Goldman, 2011) supports semi-automated alignment of various languages (including Spanish) from within Praat. Spanish models are trained on 2.9 hours of Castilian read speech.
- **Mandarin - LDC Aligner** `https://www.ldc.upenn.edu/language-resources/tools/ldc-word-aligner`

Thank You for Having Me!

Thanks to Barbara Bullock, Jacqueline Toribio, and many others for making this workshop a success!

If you have any questions or need some pointers after the workshop, feel free to email me!
wilbanks.ericw@gmail.com

References

Cuetos, F., Glez-Nosti, M., Barbón, A., and Brysbaert, M. (2011). Subtlex-esp: Spanish word frequencies based on film subtitles. *Psicológica*, 32:133–143.

Goldman, J.-P. (2011). Easyalign: an automatic phonetic alignment tool under praat. Proceedings of *Interspeech*, Firenze, Italy.

Gorman, K., Howell, J., and Wagner, M. (2011). Prosodylab-aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics*, 39(3):192–193.

Labov, W., Rosenfelder, I., and Fruehwald, J. (2013). One hundred years of sound change in philadelphia: Linear incrementation, reversal, and reanalysis. *Language*, 89(1):30–65.

Lubbers, M. and Torreira, F. (2013-2015). Praatalign: an interactive praat plug-in for performing phonetic forced alignment. `https://github.com/dopefishh/praatalign`. Version 1.7a.

Milne, P. (2014). *The variable pronunciations of word-final consonant clusters in a force aligned corpus of spoken French*. PhD thesis, University of Ottawa.

Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., and Fosler-Lussier, E. (2007). Buckeye Corpus of conversational speech (2nd release). [`www.buckeyecorpus.osu.edu`] Columbus, OH: Department of Psychology, Ohio State University (Distributor).

Schiel, F. and Draxler, C. (2003). *The production of speech corpora*. Bavarian Archive for Speech Signals.

Weide, R. L. (1994). Cmu pronouncing dictionary. `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`.

Yaeger-Dror, M., Beaudrie, A., Beaudrie, S., and Tania, G. (2004). Callfriend southern english corpus. Accessed via TalkBank.

Yuan, J. and Liberman, M. (2008). Speaker identification on the scotus corpus. Proceedings of Acoustics '08.