The Development of **FASE**: Forced Alignment System for Español and Implications for Sociolinguistic Methodologies

Eric Wilbanks

North Carolina State University

NWAV 44, Toronto
October 24, 2015

Overview of Talk

Motivation

Forced Alignment

- Over the past decade, technologies from speech recognition have begun to be utilized in phonetic research.
- Forced alignment takes as input an orthographic transcription and audio file and creates as output a time-aligned phonological (or possibly phonetic) transcription.

Benefits for Phonetics

- Manual segmentation of phones is incredibly time-consuming, at some estimates 800x real-time (Schiel and Draxler, 2003).
- Completely automated transcription/segmentation is still a work in progress (c.f. Reddy and Stanford, 2015)
- 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.

P2FA

- The mostly widely used acoustic models used for English forced aligning are part of the **Penn Phonetics Lab Forced Aligner** (Yuan and Liberman, 2008, P2FA).
- Trained on a large corpus of Supreme Court Justice oral argument recordings; Extremely robust for North American English
- These acoustic models are also adapted for use in the **Forced Alignment and Vowel Extraction** suite (Rosenfelder et al., 2011, FAVE).

Non-English

- Comparable systems for languages other than English are not yet as widely researched or utilized.
- **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, including Spanish
- **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.

Goals

- Report on the validity of utilizing messy sociolinguistic interviews to train forced alignment systems, in place of clean read speech.
- Argue for speaker adaptations by linear transforms in both training and alignment to improve alignment across a variety of recording environments and speakers.
- Demonstrate application of aligner: /d/ lenition within the corpus

Acoustic Models

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.

Dictionary Construction

- In order to carry out training and aligning, a pronunciation dictionary is needed which maps words to strings of phones.
- The dictionary was constructed from the 44 million words **SUBTLEX-ESP** corpus (Cuetos et al., 2011).
- Spanish orthography is very close to phonological representation, making conversion of words to phone sequences easy.
- 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

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)

Processes to Note

- Latin American Spanish, therefore we have no /s/-/θ/ or /l/ - /ʎ/ distinctions
- No distinction made between tonic and atonic vowels
- No distinction between high vowels and their glide allophones

Technical Specifications

Model training and application was carried out using the HTK suite (Young et al., 2006) with the following parameter values.

- WAV files downsampled to 11025hz, 11 mfcc coefficients, delta, and delta-delta extracted
- Emitting state of **sp** tied to the central state of the **silence** phone.

	States	**Gaussians**
sp	3	32
sil	5	32
others	5	16

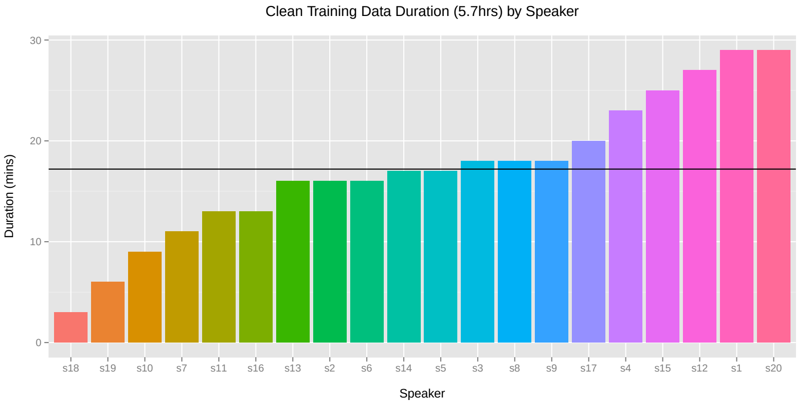
CERD

- The Corpus del Español de Raleigh-Durham (CERD) contains over 240 sociolinguistic interviews conducted in Spanish between 2008-present.
- Speakers come from a variety of regions, though most speakers (or their families) are from Mexico, Colombia, or Puerto Rico.
- Include variable experience with English, Heritage Speakers to 1-2 years in the US.

Transcription

- 20 Interviews were chosen to be orthographically transcribed and used as training data.
- Balanced for sex and age group
- Speakers were either from Mexico or of Mexican descent.
- Notably, the variety of Spanish spoken in central Mexico tends to resist elision processes typical of other varieties. Ideal for training models.
- Orthographic transcriptions were carried out by native Spanish L1 speakers

Training Data



Manual Segmentation

- Two trained Spanish phoneticians individually hand-segmented 100s of speech from a sociolinguistic interview.
- Speaker is external to the training data, young female speaker born in Mexico who moved to North Carolina at a young age.
- Differences between boundary placement/segment duration are computed between the two human transcribers and between each transcriber and the model.

Model Comparison

- **M1**: no speaker adaptation during training
- **M2**: adaptation during training; acoustic models updated via Constrained Maximum Likelihood Linear Regression (CMLLR) transforms

Why might we need Adaptation?

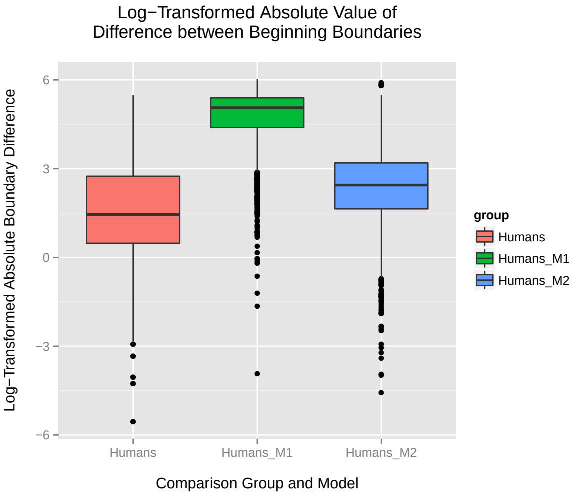
Good Quality

"El clima, ahorita está haciendo buen clima, pero sí cuando se cerca invierno,"
The weather, right now the weather's good, but yeah when winter comes,

Bad Quality

"Y su hermano menor estudia,"
And their younger brother studies,

Log Beginning



Beginning Linear Model

1. No sig. difference between HumanA-Model and HumanB-Model ($p = 0.17$)
2. **begin_diff** significantly lower in HumanA-HumanB group than in HumanA-Model and HumanB-Model ($p < 0.001$) groups.
3. M2 (adapted) has significantly lower boundary differences than Model 1; ($\beta = -3.616$, $p < 0.001$)

Linear model; dependent: absolute value of Beginning Difference, independent: Group and Model

lm(abs(begin_diff) ~ group + model, data = M1M2)

Distributions

	mean	sd	se
HumanA_HumanB	14.47ms	25.57	0.92
Humans_Model1	26.23ms	44.53	1.13
Humans_Model2	20.81ms	31.99	0.81

Descriptive Statistics of Boundary Differences by Group for Model2

Comparisons

	< **10ms**	< **20ms**
Goldman (2011)[1]	60.26%	87.11%
HumanA_HumanB	68.38%	78.66%
Humans_Model1	37.28%	63.30%
Humans_Model2	45.05%	69.41%

Percentage of Boundary Differences by Group for Models 1 and 2

[1] Human-Model; Read data rather than spontaneous

Application

/b,d,g/ Lenition - [β,ð,ɣ]

- Spanish voiced stops alternate between occlusive and approximant realizations; traditionally considered a binary distinction (Tomás, 1967)
- Recent work demonstrates it's best considered a gradient process (Lewis, 2001)
- Acoustic/Articulatory realizations conditioned by a variety of segmental, prosodic, lexical, and morphological variables

/d/ Specifically

Examining intervocalic /d/ we expect to see the most occlusion

1. After high vowels (Simonet et al., 2012)
2. Before high vowels (Ortega-LLebaría, 2004)

Additionally, preceding environment tends to have the stronger effect (Simonet et al., 2012).

Methodology

- All intervocalic /d/ tokens from the 20 training speakers extracted
- Exclusion of "De" (21% of data) leaves 1,482 tokens.
- Following Hualde et al. (2011), the difference in intensity between the following vowel and the /d/ is calculated.
- If a value is closer to 0, it indicates the /d/ is more open and less occlusive

Intensity Example

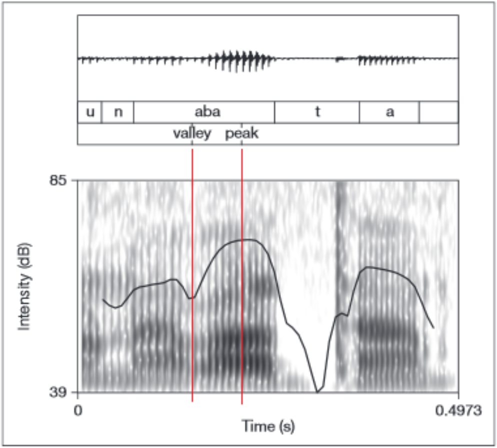
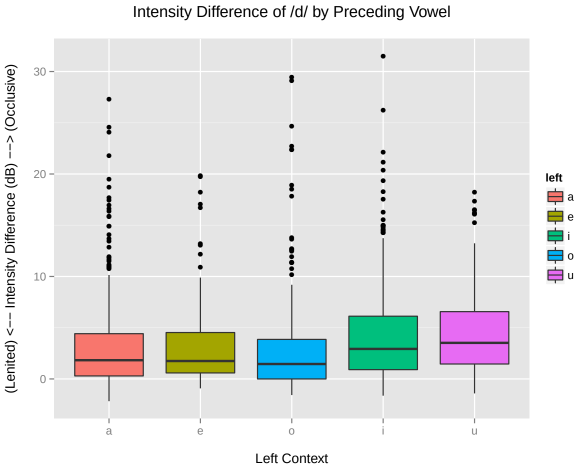
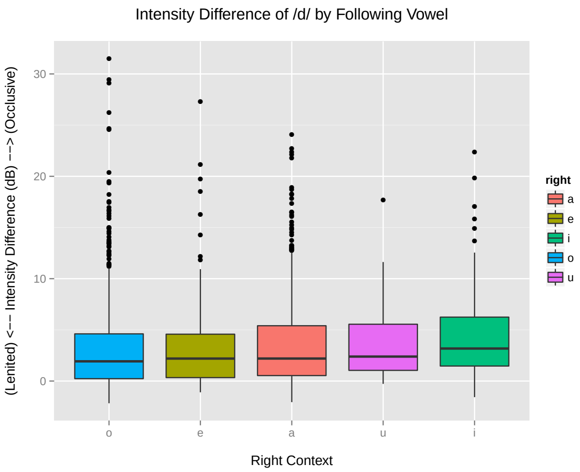


Image adapted from Carrasco et al. (2012, pp. 156)

Preceding Segment



Following Segment



Linear Mixed Model

- /d/ sig. more occlusive when preceded by /i,u/ > /a,e,o/
- /d/ sig. more occlusive when followed by /i/ > /u,a,e,o/

lmer(intensity_diff ~ left + right + (1|speaker),data = df)

Take Home Points

- Using speaker adaptation, sociolinguistic corpora make excellent training data for new forced aligners
- FASE produces excellent alignments of novel data, although not surpassing human transcription
- Using automatic alignments, well-studied internal constraints of /d/ lenition were reproduced within the corpus.

References

Carrasco, P., Hualde, J. I., and Simonet, M. (2012). Dialectal differences in spanish voiced obstruent allophony: Costa rica versus iberian spanish. *Phonetica*, 69:149–179.

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.

Hualde, J. I., Shosted, R., and Scarpace, R. (2011). Acoustics and articulation of spanish /d/ spirantizaion. In *Proceedings of the 19th International Congress on the Phonetic Sciences, Hong Kong*.

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.

Lewis, A. M. (2001). *Weakening of Intervocalic /P, T, K/ in Two Spanish Dialects: Toward the Quantification of Lenition Processes*. PhD thesis, University of Illinois at Urbana-Champaign.

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.

Ortega-LLebaría, M. (2004). Interplay between phonetic and inventory constraints in the degree of spirantization of voiced stops: comparing intervocalic /b/ and intervocalic /g/ in spanish and english. In Face, T. L., editor, *Laboratory approaches to Spanish phonology*, pages 237–253. Mouton de Gruyter, Berlin, 1 edition.

Reddy, S. and Stanford, J. N. (2015). A web application for automated dialect analysis. Proceedings of NAACL 2015.

Rosenfelder, I., Fruehwald, J., Evanini, K., and Yuan, J. (2011). Fave (forced alignment and vowel extraction) program suite. http://fave.ling.upenn.edu.

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

Simonet, M., Hualde, J. I., and Nadeu, M. (2012). Lenition of /d/ in spontaneous spanish and catalan. In *Interspeech*, pages 1416–1419.

Tomás, T. (1967). *Manual de pronunciación española*. Consejo Superior de Investigaciones Científicas. Instituto "Miguel de Cervantes." Publicaciones de la Revista de filología española. Graficas Monteverde, S.A.

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

Young, S. J., Evermann, G., Gales, M. J. F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. C. (2006). *The HTK Book, version 3.4*. Cambridge University Engineering Department, Cambridge, UK.

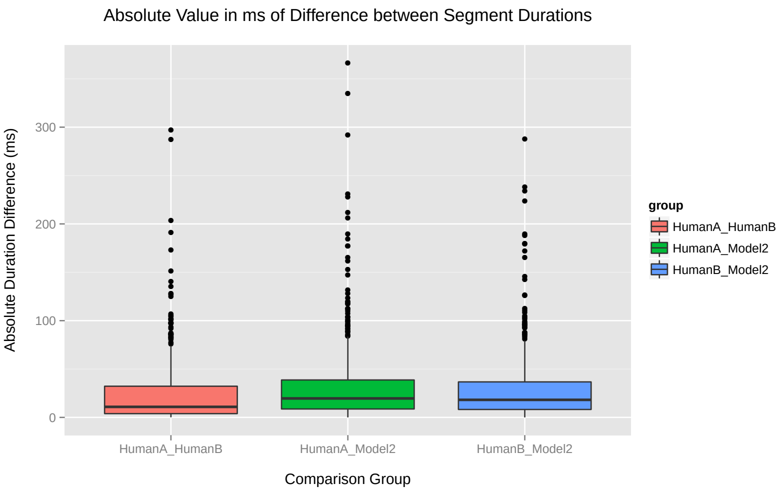
Yuan, J. and Liberman, M. (2008). Speaker identification on the scotus corpus. Proceedings of Acoustics '08.

Thank you! Gracias!

Questions, Comments, Suggestions?

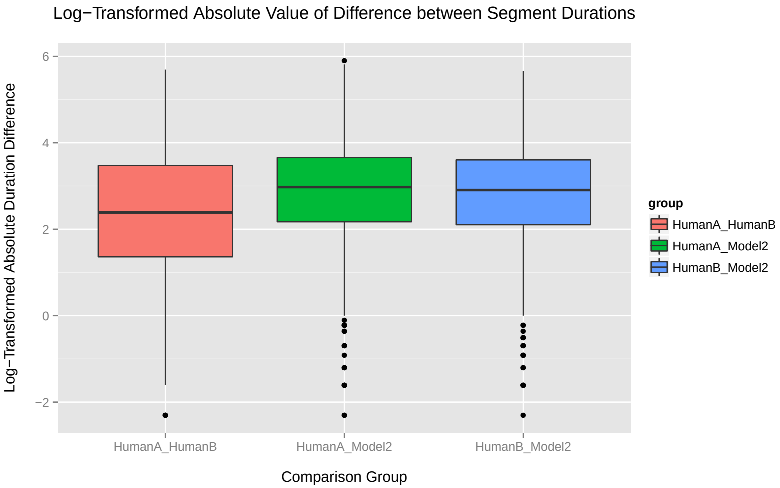
@eric_wilbanks
ewwilban@ncsu.edu
ericwilbanks.github.io

Duration

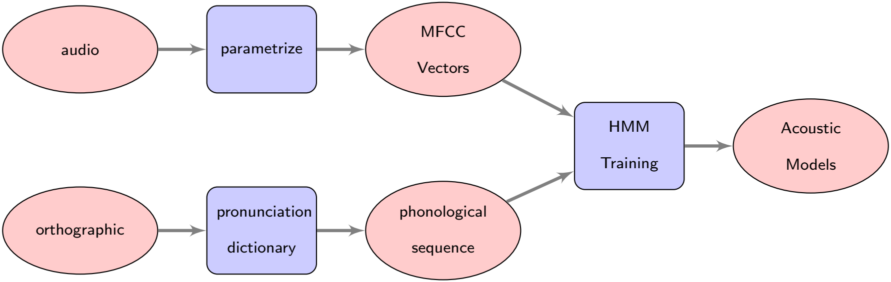


Absolute Value in ms of Difference between Segment Durations
Absolute Duration Difference (ms)
300
200
100
0
HumanA_HumanB
HumanA_Model2
HumanB_Model2
Comparison Group
group
HumanA_HumanB
HumanA_Model2
HumanB_Model2

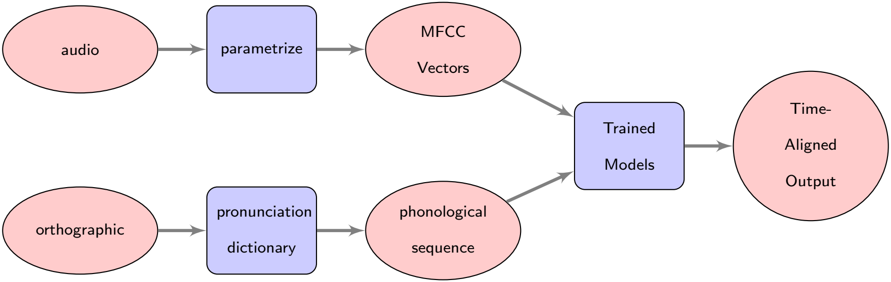
Log Duration



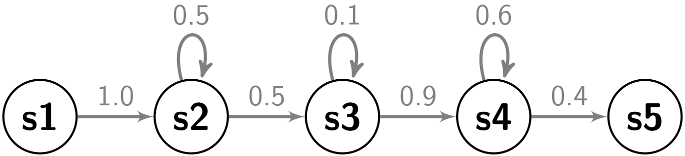
Training Acoustic Models



Generating Alignments



Left-To-Right HMM Model of Phone



Linear Mixed Model

Left	a	e	i	o	u		Right	a	e	i	o	u
a	×	0.21	-3.53	1.05	-3.68		a	×	-0.95	-2.44	-0.26	-0.61
e	-0.21	×	-2.87	0.69	-3.47		e	0.95	×	-1.09	0.79	-0.16
i	3.53	2.87	×	3.76	-1.34		i	2.44	1.09	×	2.23	0.42
o	-1.05	-0.69	-3.76	×	-3.98		o	0.26	-0.79	-2.23	×	-0.54
u	3.68	3.47	1.34	3.98	×		u	0.61	0.16	-0.42	0.54	×

Columns are Reference Levels

lmer(intensity_diff ~ left + right + (1|speaker),data = df)

Linear Mixed Model

Left	a	e	i	o	u		Right	a	e	i	o	u
a	x						a	x				
e		x					e		x			
i			x				i			x		
o				x			o				x	
u					x		u					x

Columns are Reference Levels

Green - Sig. More Lenited /d/

Red - Sig. More Occlusive /d/

lmer(intensity_diff ~ left + right + (1|speaker),data = df)