Automatic Alignment in Yoloxóchitl Mixtec documentation corpora

Christian T. DiCanio[a], Hosung Nam[a], D. H. Whalen[a,b,c]
H. Timothy Bunnell[d], Jonathan D. Amith[e], and Rey Castillo García[f]

(a) Haskins Laboratories, (b) CUNY Graduate Center, (c) Endangered Language Fund,
(d) University of Delaware, (e) Gettysburg College, (f) CIESAS-Mexico

—

dicanio@haskins.yale.edu

1/5/13

From the corpus to the description

1. Endangered language documentation projects typically involve the collection of 30+ hours of spontaneous and elicited speech. The emphasis on collecting high quality acoustic recordings means that we now have an immense amount of data for descriptive purposes.

2. The continuing vitality of these corpora depends on how accessible they remain for both community use and linguistic analysis.

3. Yet, the task of turning this data into something usable is *huge* in terms of time and resources – transcription, segmentation, verification.

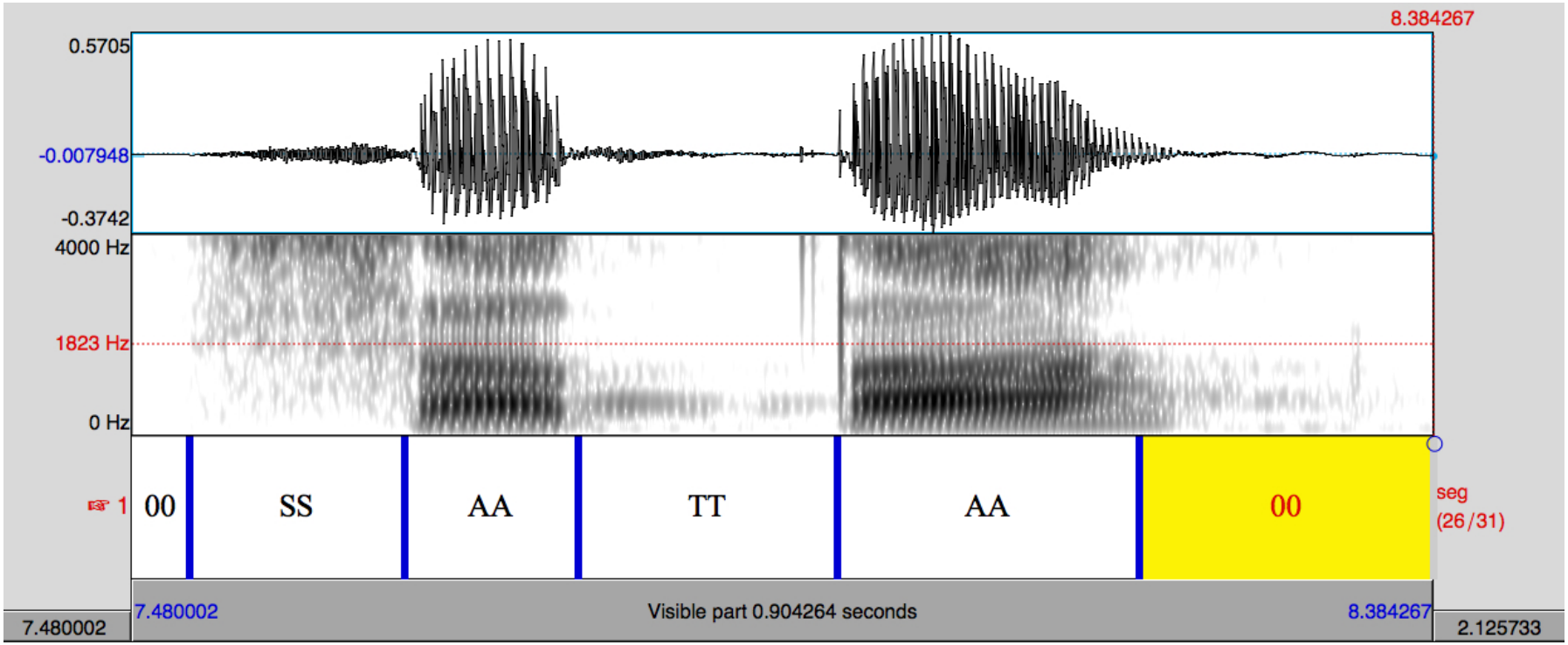
4. Are there tools to help us solve this problem? How well do they work?

Forced Alignment

- While there is no quick tool for doing transcription, *forced alignment* is one tool available to help with the task of segmenting speech.
- Using a lexicon of words, a transcription of the speech signal, and the speech signal, a forced alignment system can create a segmentation of the speech signal automatically.
- Forced aligners are language-specific. They are trained on a corpus of data from a language where segmentation by hand has already been done. These data are used to build hidden markov models for the acoustic signature of each phone. The forced alignment system then uses its internal model to predict where boundaries between phones occur.

Example from Yoloxóchitl Mixtec

Input: sound file, transcription /sata/, and lexicon containing way of coding each transcription, e.g. /sata/ = SSAATTAA



Once alignment is done, we can use scripts written for Praat to extract acoustic phonetic measures from intervals.

Research questions

1. Most systems are trained on major languages such as English (Malfrère et al., 2003; Yuan and Liberman, 2008, 2009), French (Adda-Decker and Snoeren, 2011; Malfrère et al., 2003), Spanish (Malfrère et al., 2003), Dutch (Malfrère et al., 2003), and Mandarin Chinese (Lin et al., 2005) but a few are trained on less well-studied languages like Gaelic (Ní Chasaide et al., 2006) and Xhosa (Roux and Visagie, 2007).

2. How well do existing aligners based on English work in segmenting elicited data from Mixtec?

3. How well do the same aligners work in segmenting *running* corpus speech from Mixtec?

Roadmap

1. Background on language data
2. Description of two forced alignment systems: P2FA ("Penn aligner") (Yuan and Liberman, 2008) and hm-Align (Bunnell et al., 2005)
3. Comparison of aligner performance with elicited data.
4. Examination of aligner performance with running speech data.
5. Discussion: aligner differences, potential problems/solutions.

The corpus - Yoloxóchitl Mixtec

- Yoloxóchitl Mixtec (YM) is an endangered Mixtec variant (Oto-Manguean) spoken in Guerrero, Mexico.
- Topic of large scale documentation project with 70+ hours of transcribed narratives (Amith, Castillo-García) and topic of investigation into tonal phonetics and phonology (Castillo García, 2007; DiCanio et al., 2012).
- Corpus of 261 words spoken in isolation, repeated 6 times, by 10 native speakers = 15,660 words. Collected for initial analysis of tonal phonetics. These consist of monosyllables and disyllables , e.g. /ko^{1}o^{4}/ '*snake*', /^{n}da^{1}βa^{1}/ '*wooden staff*

Mixtec phonology

- Simple segmental inventory of 16 consonants and 5 vowels (and nasal vowels), including voiceless unaspirated stops and pre-nasalized stops.
- Simple syllable structure (CV). Words are minimally bimoraic (CVV, CVCV) and maximally trimoraic (CVCVV, CVCVCV).
- Glottalization is a feature of the bimoraic foot and surfaces intervocalically and preceding sonorants, e.g. /koʔ^{1}o^{4}/ '*plate*', /^{n}daʔ^{1}βa^{1}/ '*to be turned off*
- Large tonal inventory, consisting of four level tones and five contour tones. Up to twenty tone melodies are possible on bimoraic words. The mora is the TBU (DiCanio et al., 2012).

Segmental inventory

Table 1: Yoloxochitl Mixtec Consonant Inventory

	Bilabial	Dental	Post-alveolar	Palatal	Velar	Labialized Velar	Glottal
Plosive	(p)	t			k	k^w	ʔ
Pre-nasalized plosive	(mb)	nd					
Affricate			tʃ				
Nasal	m	n					
Tap		(ɾ)					
Fricative	β	s	ʃ				
Approximant		l		j			

Table 2: Yoloxochitl Mixtec Vowel Inventory

	Front	Central	Back
Close	i, ĩ		u, ũ
Close-mid	e, ẽ		o, õ
Open		a, ã	

Alignment systems

We compared the accuracy of two alignment systems trained on English against hand-labelling on the Mixtec data.

- P2FA (Yuan and Liberman, 2008, 2009)
  - Uses GMM-based monophone-HMMs trained using the SCOTUS corpus, which consists of Supreme Court arguments.
  - CMU phone set (phonemic)

- hm-Align (Bunnell et al., 2005)
  - Uses a set of discrete monophone HMMs trained on data from the TIMIT corpus (Garofolo et al., 1993), which consists of read speech.
  - A stand-alone version of the aligner developed for the ModelTalker TTS system's voice recording program.
  - ASEL Extended English phone set (allophonic)

Phone sets and correspondences with YM

Mixtec	P2FA	hmAlign
/p/ [p]	P [p^h, p]	PP [p]
/t/ [t]	T [t^h, t, $\widehat{tʔ}$, ɾ]	TT [t]
/k/ [k]	K [k^h, k]	KK [k]
/k^w/ [k^w]	K [k^h, k]	KK [k]
/ʔ/ [ʔ]	T [t^h, t, $\widehat{tʔ}$, ɾ]	TQ [$\widehat{tʔ}$]
/^{n}d/ [^{n}d]	N [n]	NN [n]
/tʃ/ [tʃ]	CH [tʃ]	CH [tʃ]
/m/ [m]	M [m]	MM [m]
/n/ [n]	N [n]	NN [n]
/β/ [β, β̞, b]	W [w]	WW [w]
/s/ [s]	S [s]	SS [s]
/ʃ/ [ʃ]	SH [ʃ]	SH [ʃ]
/ɾ/ [ɾ]	R [ɻ, ɹ]	RR [ɻ, ɹ]

Mixtec	P2FA	hmAlign
/l/ [l]	L [l, ɬ]	LL [l, ɬ]
/j/ [j]	Y [j]	JY [j]
/i/ [i]	IY [i, ĩ]	II [i, ĩ]
/ĩ/ [ĩ]	IY [i, ĩ]	II [i, ĩ]
/e/ [e, ɛ]	EH [ɛ, ɛ̃]	EH [ɛ, ɛ̃]
/ẽ/ [ẽ, ɛ̃]	EH [ɛ, ɛ̃]	EH [ɛ, ɛ̃]
/a/ [a]	AA [ɑ, ɑ̃, a, ã]	AA [ɑ, ɑ̃, a, ã]
/ã/ [ã]	AA [ɑ, ɑ̃, a, ã]	AA [ɑ, ɑ̃, a, ã]
/o/ [o, ɔ]	AO [ɔ, ɔ̃]	AO [ɔ, ɔ̃]
/õ/ [õ, ɔ̃]	AO [ɔ, ɔ̃]	AO [ɔ, ɔ̃]
/u/ [u]	UW [u, ũ, ʉ, ʉ̃]	UW [u, ũ, ʉ, ʉ̃]
/ũ/ [ũ]	UW [u, ũ, ʉ, ʉ̃]	UW [u, ũ, ʉ, ʉ̃]

Testing method

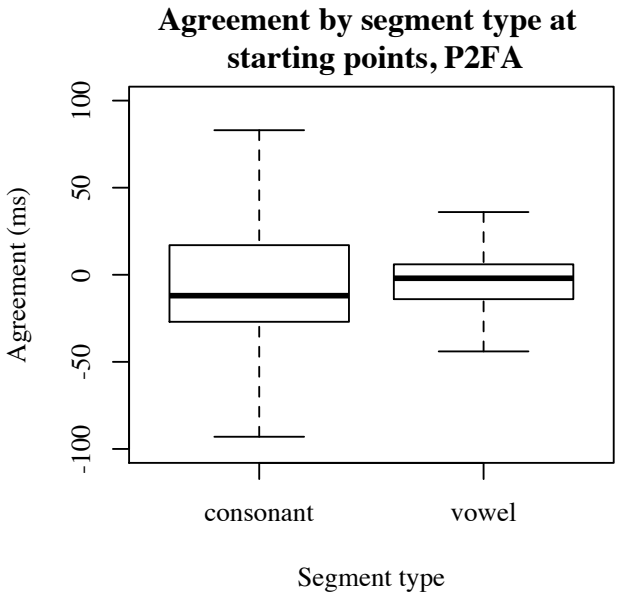
- Hand-labelling (segmentation) of YM corpus by graduate student, with correction and additional segmentation by the first author.
- Time indices were extracted using a Praat script (Boersma and Weenink, 2009) and analyzed using R (R Development Core Team, 2009). The relative differences between the segment boundaries for the hand-labeled files and the force-aligned files were compared.
- All words were coded for consonant natural class, the presence of glottalization, size (monosyllabic/disyllabic), and tone.
- Lexical items were treated as a random effect in a linear mixed effects model (lmer) and the aligner (P2FA or hm-Align) and phonological category were treated as independent variables.

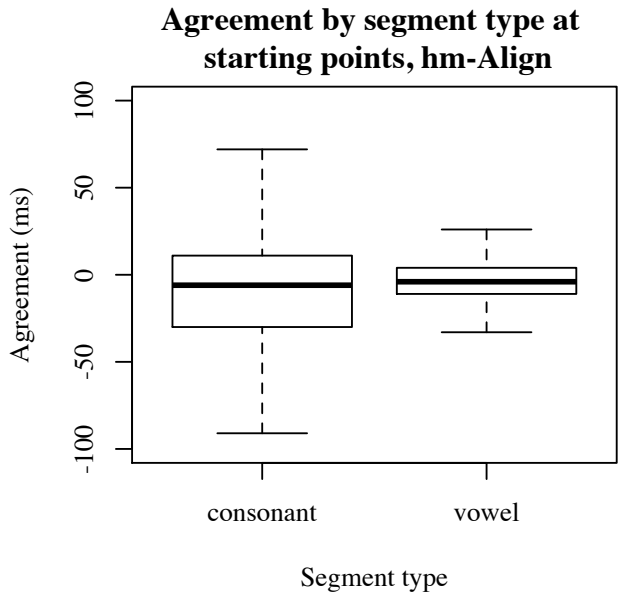
Results I: Global patterns

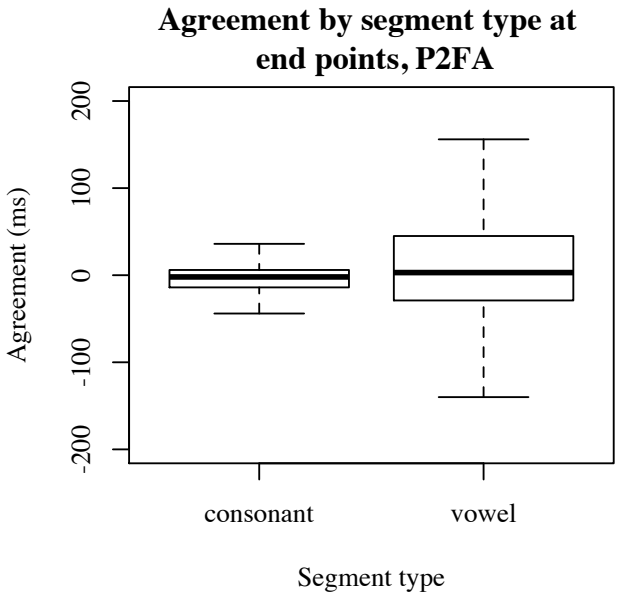
Overall, agreement for hm-Align was better than for P2FA, with a 14.8% error reduction at 20 ms. A strong effect of aligner on agreement was found, both at start points and at endpoints. Note that forced alignment estimates boundaries only to the nearest 10 ms.

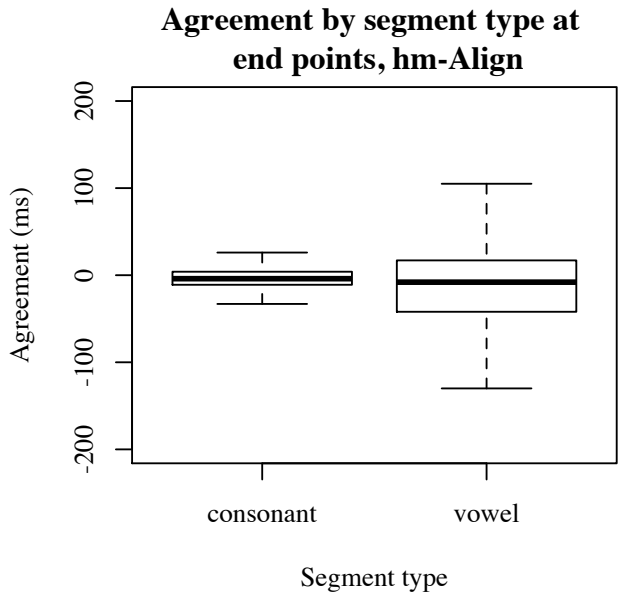
Threshold	P2FA	hm-Align
10 ms	32.3%	40.6%
20 ms	52.3%	61.4%
30 ms	65.7%	70.9%
40 ms	74.8%	81.2%
50 ms	79.6%	86.7%

Generally, agreement is between 70-90% accuracy at 20 ms (Malfrère et al., 2003), so this is low compared to aligners trained on the target language.


Agreement by segment type at starting points, P2FA
Agreement (ms)
100
50
0
-50
-100
consonant
vowel
Segment type


Agreement by segment type at starting points, hm-Align
Agreement (ms)
100
50
0
-50
-100
consonant
vowel
Segment type


Agreement by segment type at end points, P2FA
Agreement (ms)
200
100
0
-100
-200
consonant
vowel
Segment type


Agreement by segment type at end points, hm-Align
Agreement (ms)
200
100
0
-100
-200
consonant
vowel
Segment type

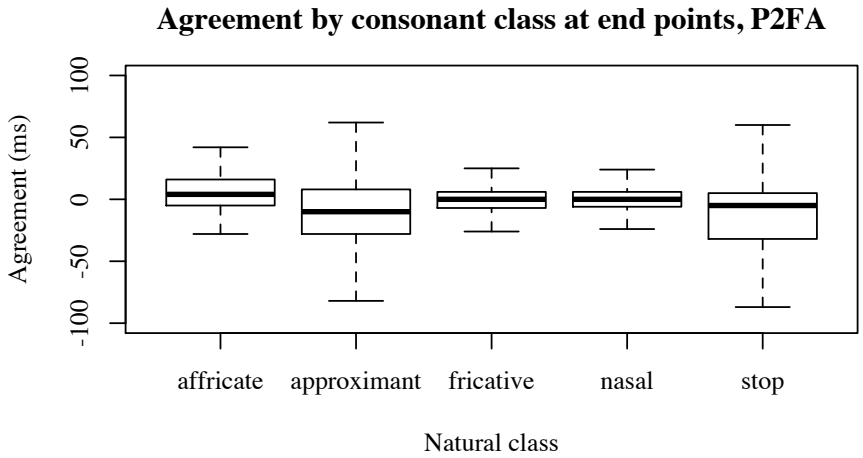
Accuracy within words and at word boundaries

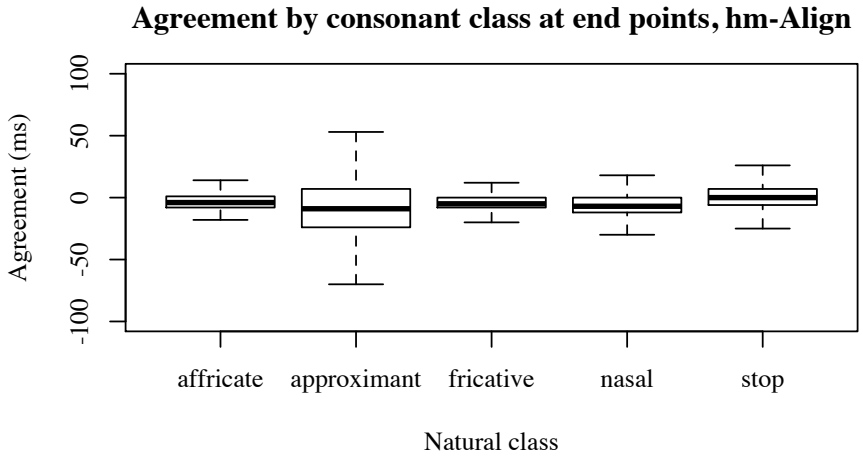
Agreement word-internally was significantly better than agreement at word boundaries. This effect was larger for hm-Align than for P2FA.

P2FA	Threshold	#-C	C-V	V-C	V-#
	10 ms	11.9%	59.4%	40.6%	48.1%
	20 ms	35.7%	70.3%	59.6%	54.7%
	30 ms	64.3%	77.6%	67.9%	63.2%
	40 ms	84.6%	81.5%	73.0%	71.7%
	50 ms	91.8%	84.3%	76.9%	77.4%
hm-Align	Threshold	#-C	C-V	V-C	V-#
	10 ms	23.2%	54.6%	46.0%	33.6%
	20 ms	43.5%	78.6%	68.9%	40.8%
	30 ms	58.1%	85.3%	77.6%	49.9%
	40 ms	82.6%	89.0%	83.2%	59.6%
	50 ms	92.2%	91.3%	87.0%	69.2%

Results II: Natural classes

Agreement was significantly better for stops with hm-Align than with P2FA.





Aligner Differences - elicited data

- Better alignment with hm-Align than with P2FA.
- Differences between aligners resulted from their training data and their phone sets. hm-Align had phones corresponding to voiceless unaspirated stops while P2FA used a "phoneme"-level phone set. The former was a better match to the YM data.
- The SCOTUS corpus (P2FA) is spontaneous speech and the TIMIT corpus (hm-Align) is read speech. One predicts that the latter would be better with more careful articulations, like those found in read speech, which is generally produced at a slower rate (Hirose and Kawanami, 2002; Laan, 1997).
- As the YM corpus consists of words in citation, hm-Align was better suited to the data. For instance, P2FA inserted many mistaken pauses in the corpus. hm-Align did not.

- The process of corpus segmentation can be aided by forced aligners trained on more common languages, but it is necessary to find an aligner with a phone set which closely matches the language's phonological system. The training corpus type is also important.

- Manual correction of existing errors can be targeted, e.g. fricative and affricate agreement was 99% at 20 ms but the worst for approximants; word boundary transitions were aligned worse than word-internal transitions.

- Phonetic description of certain characteristics may not require precise boundaries, e.g. vowel formants, fricative spectra.

Running speech data

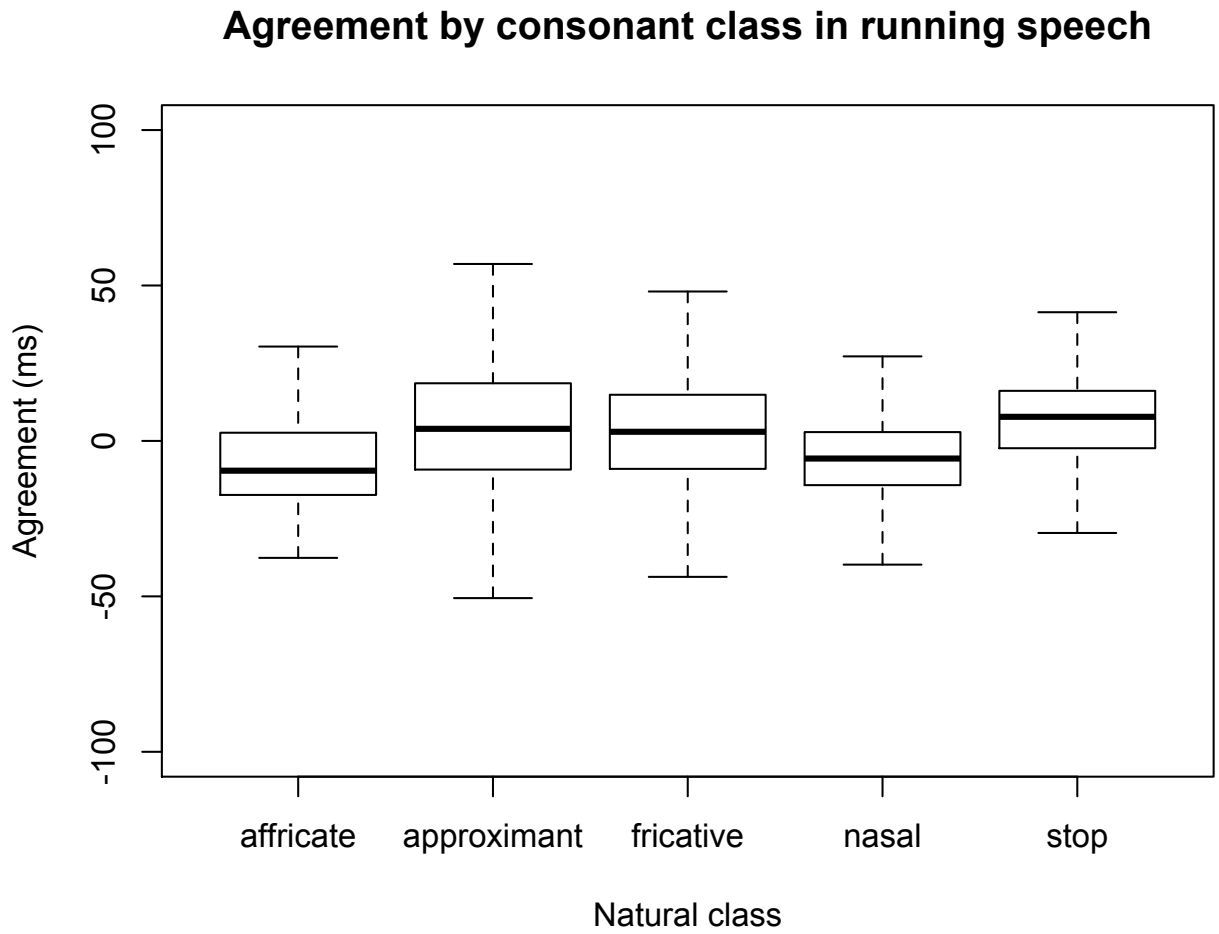
- Given that word-internal transitions are aligned better than word boundaries, one predicts forced alignment to work better for running speech data than for elicited single word utterances. This was examined here.

- A 17 minute narrative, *Adventures of the rabbit*, spoken by a 56 year old Mixtec male, was used.

- Narrative was broken into utterance-sized chunks according to time codes marked in Transcriber. These were all segmented by hand, which took roughly 22 hours (1 minute running speech = 80 minutes of segmentation).

Results

Alignment generally better than for words in isolation.



Agreement level

Approximately 18% more of the data falls within the threshold in running speech. Even though segments are shorter in running speech, this is a significant improvement.

	Elicited Speech		Running Speech
Threshold	hm-Align	P2FA	P2FA
10 ms.	40.6%	32.3%	41.3%
20 ms.	61.4%	52.3%	70.1%
30 ms.	70.9%	65.7%	83.6%
40 ms.	81.2%	74.8%	89.1%
50 ms.	86.7%	79.6%	91.5%

Issues dealing with running corpus data

1. The transcription of the corpus most likely reflects the phonemic inventory found in "careful" speech. Most texts/narratives are not careful.

2. The best transcription has surface phonetic accuracy. Unfortunately, this is often not the transcription favored by field linguists. Usually the "transcription" is a practical orthography (so not a transcription) which maintains morphological and lexical distinctiveness.

3. Language-specific reductions often occur in function words. These are **not predictable**, so they should be transcribed. Often, the full form is included in the transcription, e.g. /a^{1}tʃi^{1}/ '*before*' > [tʃi^{1}], /sa^{3}kã^{4}/ '*so*' > [sã] or [hã] or [ã].

Discussion II

- Overall, P2FA does well at forced alignment of Mixtec running speech. Considering that this aligner was built on more spontaneous speech, it may work well for this type of data.
- hm-Align has not yet been tested on running speech, but the prediction here is that it might not work as well as P2FA.
- Forced alignment is a useful tool for segmentation of elicited and corpus speech, even when one uses an aligner which is not based on the analyzed language.

Conclusions

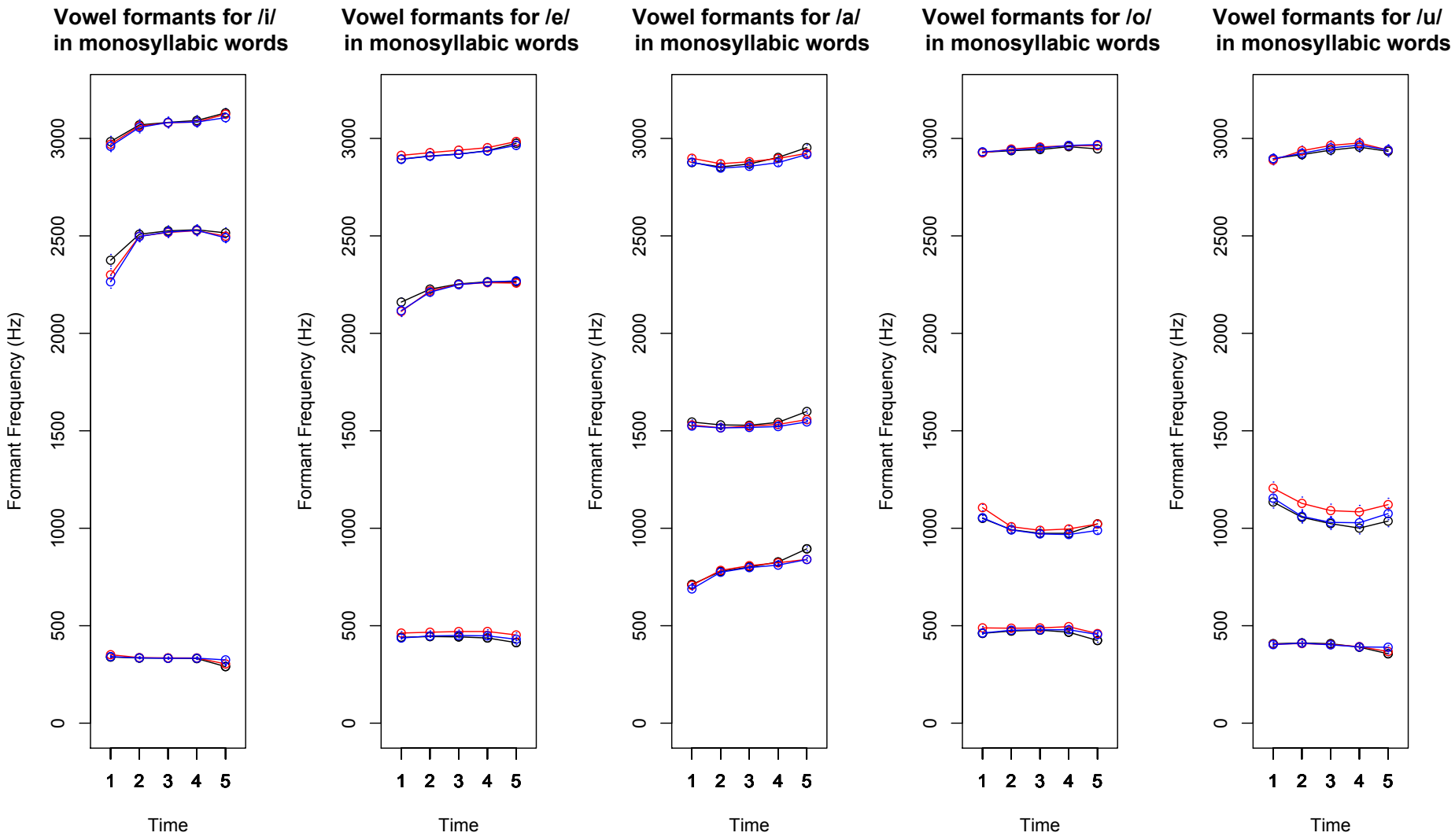
- Even though forced alignment is imperfect, the manual correction of the aligned data may take substantially less time than hand-labelling.
- If we can reduce the time frame from 80x the duration of the corpus data to only 5-10x, then this is a substantial improvement.
- The comparison of the aligners shows that it is important to consider both the phone set and the corpus data on which the aligner was built when choosing an aligner.
- Yet, the work on corpus data shows that is equally important to code and transcribe one's data appropriately so to improve the accuracy of forced alignment.

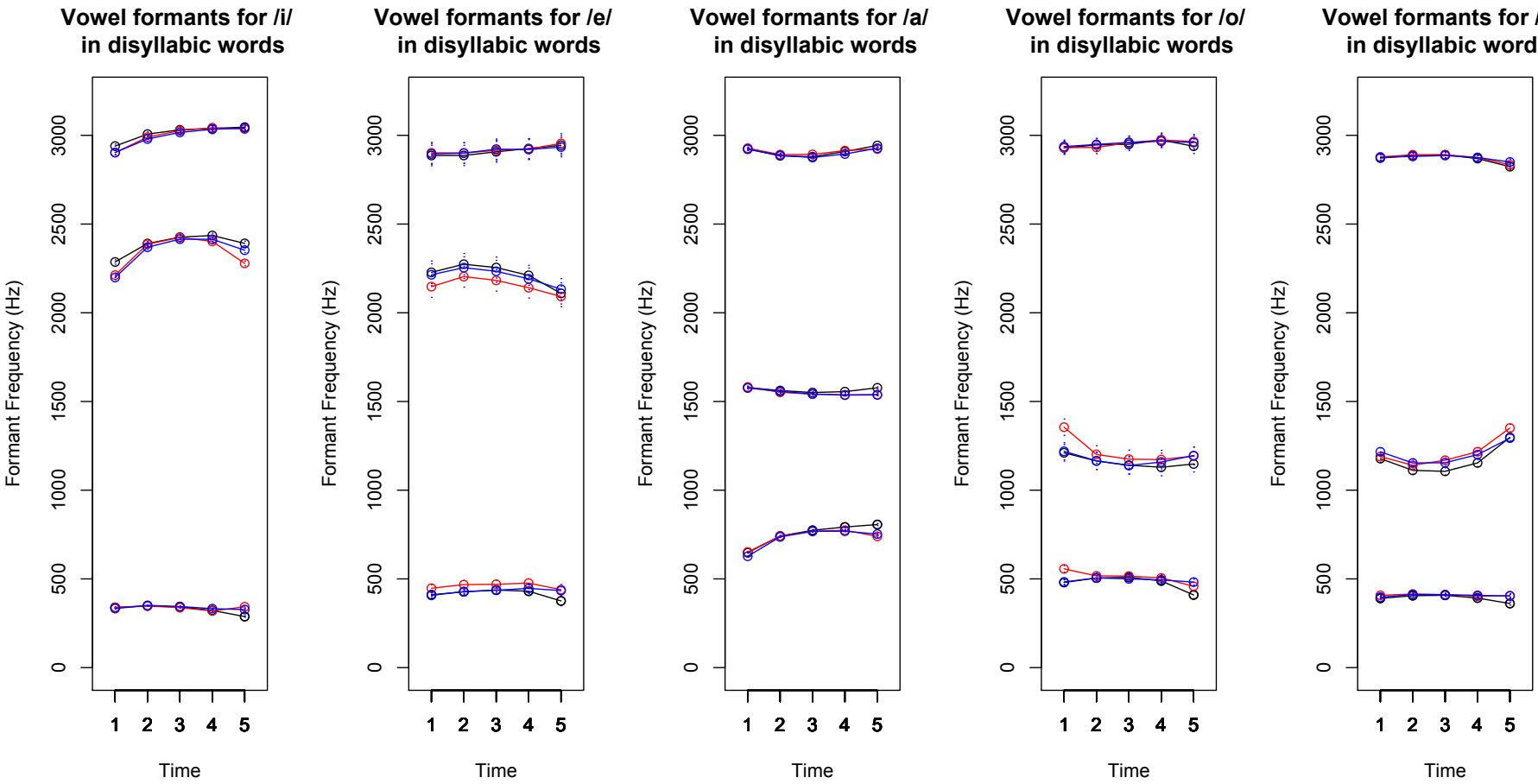
Future directions

- Assessing how much alignment accuracy matters in extracting phonetic measures (preliminary data shows promising results).
- Application of existing aligners to more spontaneous YM speech (corpus of 80+ hours transcribed texts).
- Testing TTS system based on YM (with Timothy Bunnell), which will train a language-specific forced alignment system (hm-Align).

Thank you!

Measuring vowel formants in Mixtec

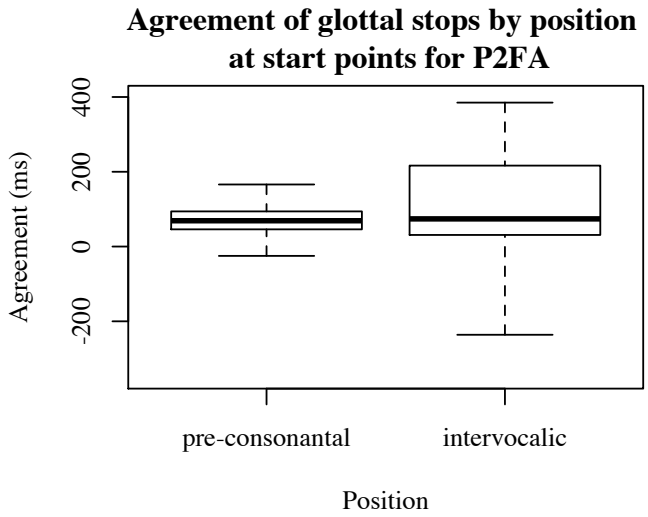


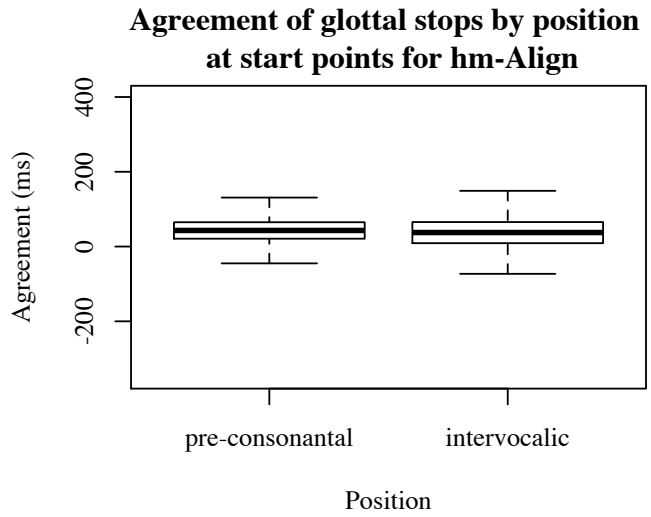

Vowel formants for /i/ in disyllabic words
Vowel formants for /e/ in disyllabic words
Vowel formants for /a/ in disyllabic words
Vowel formants for /o/ in disyllabic words
Vowel formants for /u/ in disyllabic words
Formant Frequency (Hz)
0
500
1000
1500
2000
2500
3000
Time
1
2
3
4
5

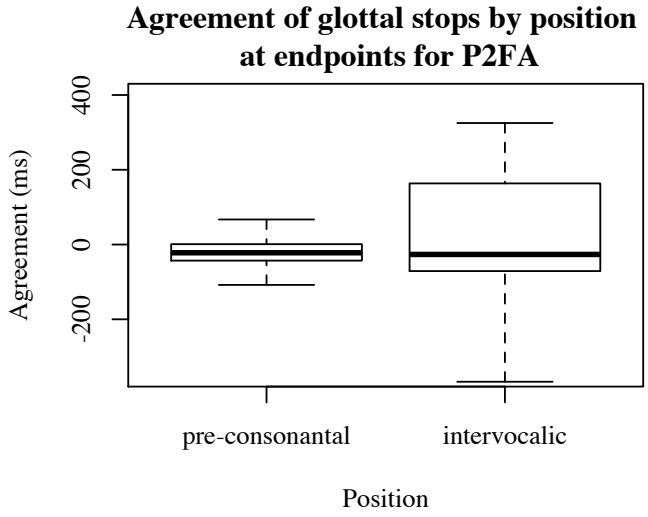
Results: Glottalization

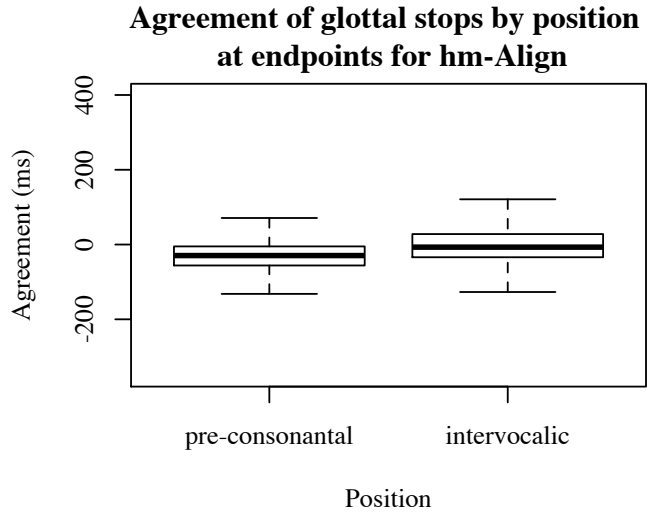
- Recall that glottalization occurs in two contexts in YM.
- The first context closely resembles glottalization occurring with coda /t/ in English (Huffman, 2005), e.g. /ʃaʔ^{4}ni^{24}/ '*killer (adj.)*' vs. /'tʃe͡tʔni/ '*chutney*'.
- The second context, intervocalic glottalization, is generally not found in American English (though there are British dialects which permit /t/ glottalization here, c.f. Foulkes and Docherty (2006)).
- Since YM glottalization resembles English glottalization in the first context, one anticipates that the alignment systems will do better here. However, note that hm-Align has a phone "TQ" specifically trained on the glottalized /t/ allophone.

Agreement for YM glottalization









Adda-Decker, M. and Snoeren, N. D. (2011). Quantifying temporal speech reduction in French using forced speech alignment. *Journal of Phonetics*, 39:261–270.

Boersma, P. and Weenink, D. (2009). Praat: doing phonetics by computer [computer program]. www.praat.org.

Bunnell, H. T., Pennington, C., Yarrington, D., and Gray, J. (2005). Automatic personal synthetic voice construction. In *INTERSPEECH-2005*, pages 89–92.

Castillo García, R. (2007). Descripción fonológica, segmental, y tonal del Mixteco de Yoloxóchitl, Guerrero. Master's thesis, Centro de Investigaciones y Estudios Superiores en Antropología Social (CIESAS), México, D.F.

DiCanio, C., Amith, J., and Castillo García, R. (2012). Phonetic alignment in Yoloxóchitl Mixtec tone. Talk Presented at the Society for the Study of the Indigenous Languages of the Americas Annual Meeting.

Foulkes, P. and Docherty, G. J. (2006). The social life of phonetics and phonology. *Journal of Phonetics*, 34:409–438.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., and Zue, V. (1993). *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Linguistic Data Consortium, Philadelphia.

Hirose, K. and Kawanami, H. (2002). Temporal rate change of dialogue speech in prosodic units as compared to read speech. *Speech Communication*, 36:97–111.

Huffman, M. K. (2005). Segmental and prosodic effects on coda glottalization. *Journal of Phonetics*, 33:335–362.

Laan, G. P. M. (1997). The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style. *Speech Communication*, 22:43–65.

Lin, C.-Y., Roger Jang, J.-S., and Chen, K.-T. (2005). Automatic segmentation and labeling for Mandarin Chinese speech corpora for concatenation-based TTS. *Computational Linguistics and Chinese Language Processing*, 10(2):145–166.

Malfrère, F., Deroo, O., Dutoit, T., and Ris, C. (2003). Phonetic alignment: speech synthesis-based vs. Viterbi-based. *Speech Communication*, 40:503–515.

Ní Chasaide, A., Wogan, J., Ó Raghaillaigh, B., Ní Bhriain, A., Zoerner, E., Berthelsen, H., and Gobl, C. (2006). Speech technology for minority languages: the case of Irish (Gaelic). In *INTERSPEECH-2006*, pages 181–184.

R Development Core Team, Vienna, A. (2009). R: A language and environment for statistical computing [computer program]. http://www.R-project.org, R Foundation for Statistical Computing.

Roux, J. C. and Visagie, A. S. (2007). Data-driven approach to rapid prototyping Xhosa speech synthesis. In *Proceedings of the 6^{th} ISCA Workshop on Speech Synthesis*, pages 143–147.

Yuan, J. and Liberman, M. (2008). Speaker identification on the SCOTUS corpus. In *Proceedings of Acoustics - 2008*.

Yuan, J. and Liberman, M. (2009). Investigating /l/ variation in English through forced alignment. In *Interspeech - 2009*, pages 2215–2218.