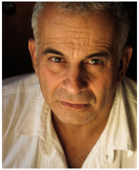
Phonetic variation and the construction of a Mixtec spoken language corpus

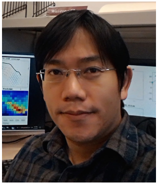
Christian DiCanio
cdicanio@buffalo.edu

Department of Linguistics
University at Buffalo

11/10/17

Acknowledgements



Doug Whalen | Jonathan Amith | Hosung Nam | Wei Rong Chen

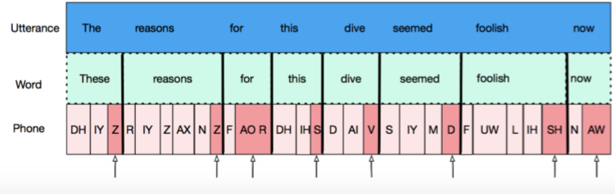
Joshua Benn (UB), Jason Lilley (Delaware), Rey Castillo García (SEP/Mexico)

National Science Foundation
WHERE DISCOVERIES BEGIN

NSF Grant No. 0966411 to Haskins Laboratories
NSF Grant No. 1603323 to the University at Buffalo

The problem: fieldwork → spoken language corpus

$\downarrow$



The documentation framework

The typical framework for language documentation involves audio/video recording, linguistic description, and transcription.

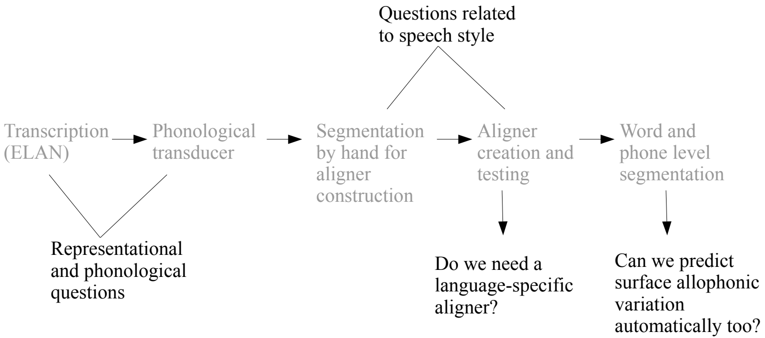
A documentation project where a team has transcribed and archived 30-40 hours of recordings is considered “complete.”

Yet in terms of speech corpus development, this reflects an early stage.

Transcription (ELAN) → Phonological transducer → Segmentation by hand for aligner construction → Aligner creation and testing → Word and phone level segmentation

Issues which arise along the way

Speech corpus development from endangered language documentation is complex and time-consuming, but research questions in speech production arise naturally in the process.



Roadmap

Issues which arise in aligner development:

1. Can we use an existing forced aligner to align the corpus? Which one?
2. Does speech style influence vowel production?

Issues which arise comparing transducer and aligner output to the speech signal:

3. Why is there so much variation in obstruent production?
4. Can we predict this in some way?

End goal: A multi-layered speech corpus that is prosodically-annotated.

The Yoloxóchitl Mixtec corpus

- Otomanguean, spoken in Guerrero, Mexico (∼2500 speakers).
- 120 hours of transcribed personal narratives, stories, and folklore; 30 speakers (Amith & Castillo García, 2009 – present).
- Phonological/phonetic fieldwork (Castillo García (2007), DiCanio et al. (2014), DiCanio (submitted a, b), Palancar et al. (2016)).

Segmental phonology

(DiCanio et al, submitted b)

	Bilabial	Dental	Alveolar	Post-alveolar	Palatal	Velar	Labialized Velar
Plosive	p	t̪				k	kʷ
Nasal	m		n				
Post-stopped nasal	mᵇ		nᵈ			ŋᶢ	
Tap			ɾ				
Affricate				tʃ			
Fricative	β	s̪		ʃ			
Approximant			l		j		

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

- All roots are minimally composed of bimoraic feet, consisting of either monosyllabic stems with long vowels (CVV) or disyllabic stems with shorter vowels (CVCV) (Castillo García, 2007). No codas.
- Glottalization occurs between vowels or before sonorants, e.g. /yaʔ^{4}a^{1}/ '*grey*', /saʔ^{3}ma^{4}/ '*cloth to wrap tortillas*'
- Final syllables are prominent.
  - Nasal vowels only occur on stem-final syllables.
  - 9 tones on stem-final syllables, but only 5 on non-final syllables.
  - Restricted vowel contrasts on non-final syllables.
  - Final syllable lengthening

Morphology	'to break' (tr)	'hang' (tr)	'to change' (intr)	'to peel' (tr)	'to get wet'
Stem	ta^{3}ʔβi^{4}	tʃi^{3}kũ^{2}	na^{1}ma^{3}	kwi^{1}i^{4}	tʃi^{3}i^{3}
NEG	ta^{14}ʔβi^{4}	tʃi^{14}kũ^{2}	na^{14}ma^{3}	kwi^{14}i^{14}	tʃi^{14}i^{3}
COMP	ta^{13}ʔβi^{4}	tʃi^{13}kũ^{2}	na^{13}ma^{3}	kwi^{1}i^{4}	tʃi^{13}i^{3}
INCOMP	ta^{4}ʔβi^{4}	tʃi^{4}kũ^{2}	na^{4}ma^{13}	kwi^{4}i^{14}	tʃi^{4}i^{4}
1S	ta^{3}ʔβi^{42}	tʃi^{3}kũ^{2}=ju^{1}	na^{1}ma^{32}	kwi^{1}i^{42}	tʃi^{3}i^{2}

Disyllabic words in YM

Twenty-six tonal melodies, including one minimal enneadecuplet (19 words).

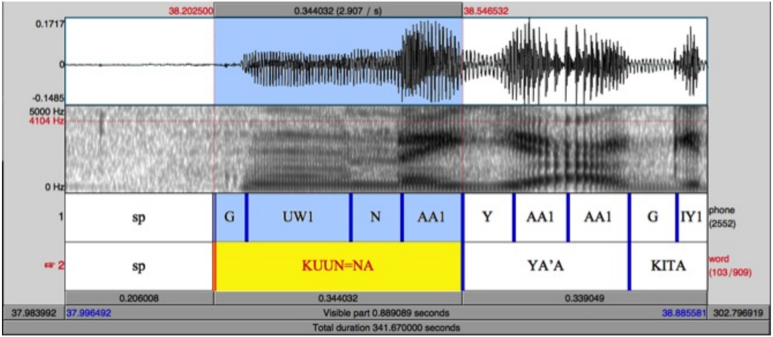
Melody	Word	Gloss	Melody	Word	Gloss
1.1	$ta^{1}ma^{1}$	*without appetite*	4.13	$na^{4}ma^{13}$	*is changing*
1.3	$na^{1}ma^{3}$	*to change (intr)*	4.14	$nda^{4}ta^{14}$	*is splitting up*
1.4	$na^{1}ma^{4}$	*soap*	4.24	$ya^{4}ma^{24}$	*Amuzgo person*
1.32	$na^{1}ma^{32}$	*I will change myself*	4.42	$na^{4}ma^{42}$	*I often pile rocks*
1.42	$na^{1}ma^{42}$	*my soap*	13.2	$hi^{13}ni^{2}$	*has seen*
3.2	$na^{3}ma^{2}$	*wall*	13.3	$na^{13}na^{3}$	*has photographed oneself*
3.3	$na^{3}ma^{3}$	*to change (tr)*	13.4	$na^{13}ma^{4}$	*has piled rocks*
3.4	$na^{3}ma^{4}$	*sprout*	14.2	$na^{14}ma^{2}$	*I will not change*
3.42	$na^{3}ma^{42}$	*I will pile rocks*	14.3	$na^{14}ma^{3}$	*to not change*
4.1	$ka^{4}nda^{1}$	*is moving (intr)*	14.4	$na^{14}ma^{4}$	*to not pile rocks*
4.2	$na^{4}ma^{2}$	*I am changing*	14.13	$na^{14}ma^{13}$	*to not change oneself*
4.3	$na^{4}ma^{3}$	*it is changing*	14.14	$nda^{14}ta^{14}$	*to not split up*
4.4	$na^{4}ma^{4}$	*is piling rocks*	14.42	$na^{14}ma^{42}$	*I will not pile rocks*

(Why a phonetician working on tone/prosody is interested in YM.)

Forced alignment

A byproduct of an acoustic model in automatic speech recognition (ASR) system, where an acoustic model is a statistical pattern classifier.

(Adda-Decker and Snoeren, 2011; Gorman et al., 2011; Malfrère et al., 2003; Yuan and Liberman, 2009)



Testing existing aligners on YM
(DiCanio et al., 2013)

What if we tried to use a forced aligner, trained on English, on YM speech to do the job? Which aligners work better?

P2FA = "the Penn aligner" (Yuan and Liberman, 2008, 2009)

- Trained using the SCOTUS corpus.
- CMU phone set (phonemic)

hm-Align (Bunnell et al., 2005)

- Trained on data from the TIMIT corpus, which consists of read speech (Garofolo et al., 1993).
- ASEL Extended English phone set (allophonic)

Phone sets and correspondences

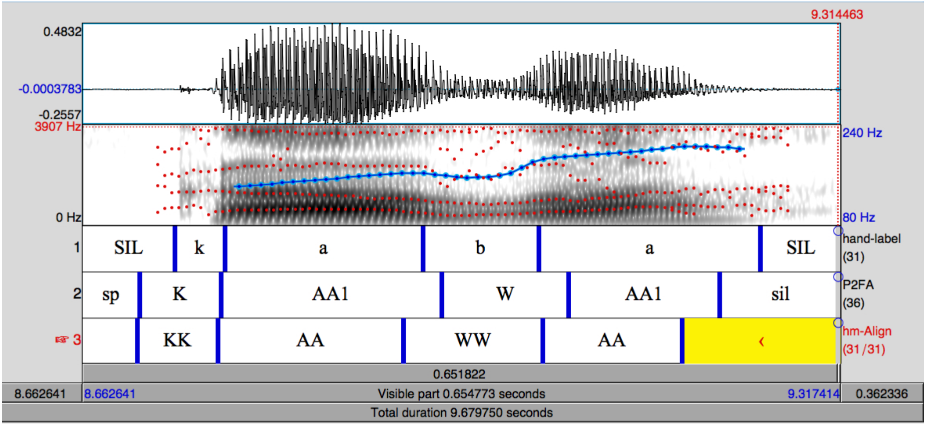
Coding for vowels/consonants, e.g. IY0 = [i] without stress, IY1 = [i] with stress; M = [m], N = [n], etc.

Mixtec	P2FA	hmAlign
/p/ [p]	P [pʰ, p]	PP [p]
/t/ [t]	T [tʰ, t, t͡ʔ, ɾ]	TT [t]
/k/ [k]	K [kʰ, k]	KK [k]
/kʷ/ [kʷ]	K [kʰ, k]	KK [k]
/ʔ/ [ʔ]	T [tʰ, t, t͡ʔ, ɾ]	TQ [t͡ʔ]

Methods

- Corpus of 261 words spoken in isolation, repeated 6 times, by 10 native speakers = 15,660 words; hand-segmented.
- These consist of monosyllables and disyllables, e.g. /ko^{1}o^{4}/ 'snake', /n^{d}a^{1}βa^{1}/ 'wooden staff'.
- Compared hand-labelled segmentation to that from forced aligners.
- Distance between boundaries compared using scripts written for Praat (Boersma and Weenink, 2016).

Example


9.314463
0.4832
-0.0003783
-0.2557
3907 Hz
0 Hz
240 Hz
80 Hz
1
SIL
k
a
b
a
SIL
hand-label
(31)
2
sp
K
AA1
W
AA1
sil
P2FA
(36)
3
KK
AA
WW
AA
hm-Align
(31/31)
0.651822
8.662641
8.662641
Visible part 0.654773 seconds
9.317414
0.362336
Total duration 9.679750 seconds

Results: general

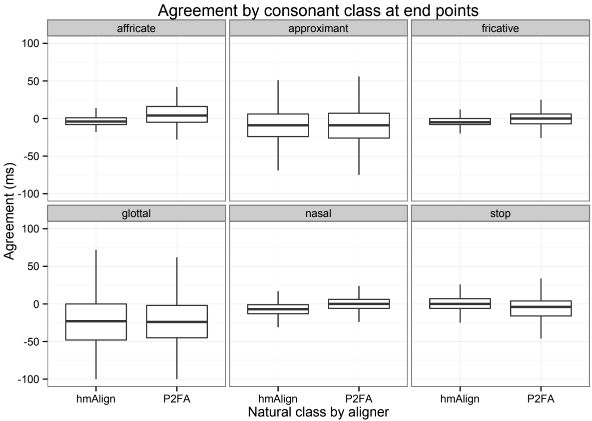
Agreement is better with hm-Align than with P2FA.

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% accurate within 20 ms (Malfrère et al., 2003). So, this is slightly less than ideal.

Results by consonant type

- Fricatives [s, ʃ, h] and nasals [m, n] are aligned well.
- Better alignment with hm-Align for stops [p, t, k] and affricates [tʃ].



Discussion

- Better alignment with hm-Align than with P2FA.
- Differences in alignment resulted from training data and phone sets.
- The SCOTUS corpus (P2FA) is spontaneous speech and the TIMIT corpus (hm-Align) is read speech (more similar to elicited Mixtec speech). The speech style used in the recordings matters!
- hm-Align phone set had voiceless unaspirated stops and a glottal stop, allowing a better match to Mixtec phonetics than P2FA's.

Segmentation in running speech

- Word-internal transitions are aligned better than word boundaries.
- Predicts forced alignment to work better for running speech data than for elicited, single word utterances.
- A 17 minute narrative, *Adventures of the rabbit*, spoken by a 56 year old Mixtec male. Segmented by hand, which took roughly 22 hours (1 minute running speech = 80 minutes of segmentation).
- Investigated only P2FA performance here as we could not retrain hm-Align on the running speech.

Better alignment!

Approximately 18% more of the data falls within the 20 ms 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%

Does speech style matter? How much?

We know that style seems to be relevant for alignment purposes, what about for generalizing about speech production?

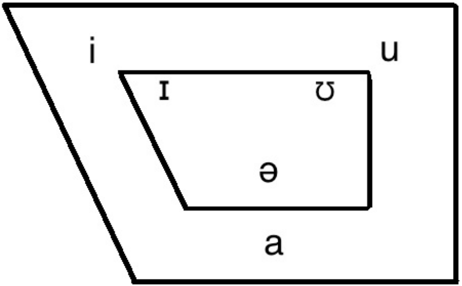
> *"From a phonetician's point of view there is no point in making lengthy recordings of folk tales, or songs that people want to sing. Such recordings can seldom be used for an analysis of the major phonetic characteristics of a language, except in a qualitative way. You need sounds that have all been produced in the same way so that their features can be compared." (Ladefoged, 2003, p.9).*

Elicited vs. spontaneous vowel production
(DiCanio et al., 2015)

- To what extent are vowels produced in a spontaneous corpus of speech similar to those produced in careful, elicited speech?
- Are patterns of vowel reduction in running speech simply a result of durational changes across register?
- Is reduction so great in spontaneous speech that one can not extract useful phonetic data? Does spontaneous speech look like elicited speech?

Vowel undershoot

Given a sufficiently short duration, the speech articulators may fail to reach an ideal vowel target, resulting in vowel *undershoot* (Lindblom, 1963, 1983, 1990; Meunier and Espesser, 2011). The more typical, reduced vowels approach a schwa-like vowel closer to the center of one's vowel space (Moon and Lindblom, 1994).



Predictions from the literature

Does speech style influence vowel production?

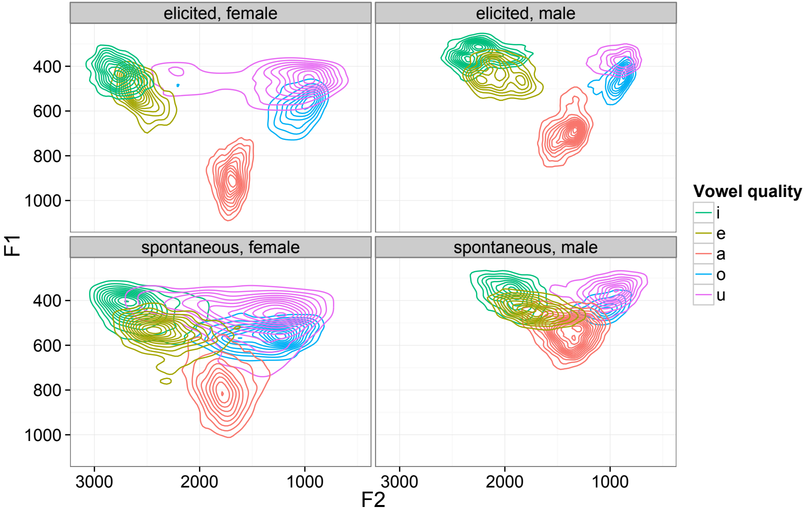
- Vowel reduction across styles is asymmetrical for back vowels (Keating and Huffman, 1984).
- Vowel reduction across styles affects peripheral vowels most (Koopmans van Beinum, 1980).
- Duration is not the only factor accounting for differences (Moon and Lindblom, 1994).

Methods

- Same elicited data used before, but now compared with 2 hrs spontaneous speech from the same speakers.
- Used P2FA to produce initial alignment of spontaneous speech data, but corrected misalignments by hand.
- Examined vowel formant data at three time points across the vowel: initial, medial, final using a script written for Praat (Boersma and Weenink, 2016). 22,167 elicited vowels and 16,219 spontaneous vowels were compared.

- Two dependent variables: intra-vowel variability and vowel dispersion, both converted to z-scores for statistical analyses.
- Four independent variables: speech style (elicited vs. spontaneous), duration, vowel (i, e, a, o, u), and sex. Contextual factors (preceding consonant place of articulation) were examined separately.
- Linear mixed effects models with random effects used (Baayen, 2008), which allow for a combination of continuous and discrete predictors, by-subject and by-item random effects, and do not require design balance.
- Model evaluation based on the Satterthwaite method to approximate for degrees of freedom, via the *lmerTest* (Kuznetsova et al., 2013) in R (R Development Core Team, 2013).

Results: Vowels in YM



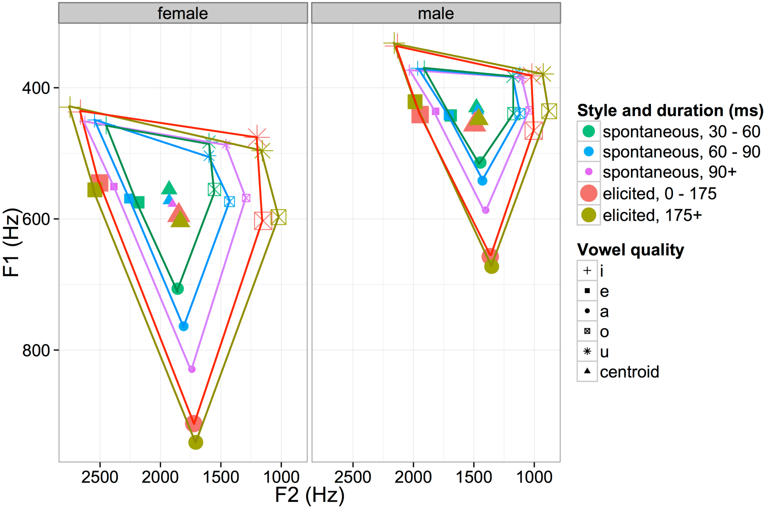
- Vowels are reduced in both space in time in spontaneous speech compared to elicited speech. Stronger effect of style on F2 dispersion than F1 dispersion.

- Females use a larger range of acoustic values in their vowel spaces; jaw opening.

- Major differences in vowel duration with style. Vowels in elicited speech were 219 ms on average, but vowels in spontaneous speech were 92 ms, a ratio of 1:2.4.

- This durational difference strongly contributed to the overall dispersion of the vowel space as a function of style. Shorter vowels were more centralized than longer ones, regardless of speech style. However, style still emerged, independently, as a significant factor to vowel variability.

Duration and the many "vowel spaces"



Discussion

Speech style involves a deformation of the vowel space which is not capturable via a single transformation.

Duration is a major contributor to such differences; a similar vowel space is observed in naturally-occurring longer duration vowels in spontaneous speech as to vowels in elicited speech.

A study on an Arapaho speech corpus found comparable durational and length-induced patterns of undershoot for long and short vowels (DiCanio and Whalen, 2015).

Though different, spontaneous speech corpora (with folktales, narratives, etc) show similar patterns to those containing elicited speech.

Building a better pronunciation dictionary for spontaneous speech

- The transcription of the corpus most likely reflects the phonemic inventory found in "careful" speech. Most texts/narratives are not careful.
- The best transcription for forced alignment should match the phonetics.
- Unfortunately, this is often not the transcription favored by field linguists. Usually the "transcription" is a practical orthography which maintains morphological and lexical distinctiveness.

Improving alignment via a phonological transducer

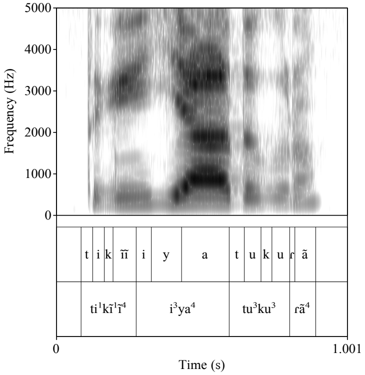
How do we tell an aligner that ki'^{3}in^3 = on^4 /kĩʔ^{3}ĩ^3 = õ^4/ 'you take' is pronounced [kʲõʔ^{3}õ^4]?

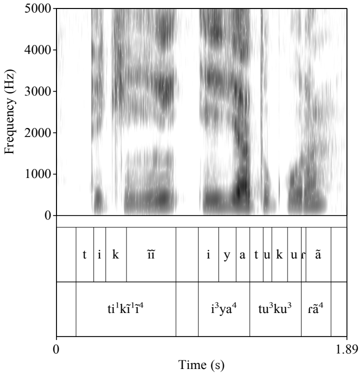
For words where there is regular pattern, we can create phonological rules that we apply to the transcription to give us something more phonetic.

1. ki'^{3}in^3 = on^4	Input
2. ki'^{3}on^4	Vowel replacement/harmony
3. ki'^{3}on^4	Replace all preceding vowels if [-high] else [+high] –> glide
4. kyo'^{3}on^4	Output

Consonant variability

But there are mismatches between the input and output despite one's best attempts at producing a transducer: '*...the sour tamale again, then.*'
[ti^{1}ɣĩ^{1}ĩ^{4} jːa^{4} du^{3}ɣu^{3} ɾã^{4}] (left) vs. [ti^{1}kĩ^{1}ĩ^{4} i^{3}ja^{4} tu^{3}ku^{3} ɾã^{4}] (right)

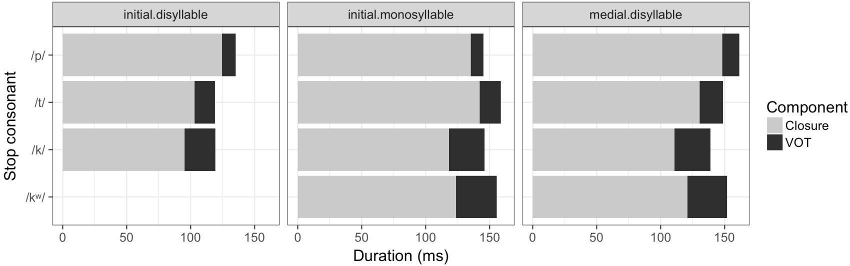




Variable obstruent lenition

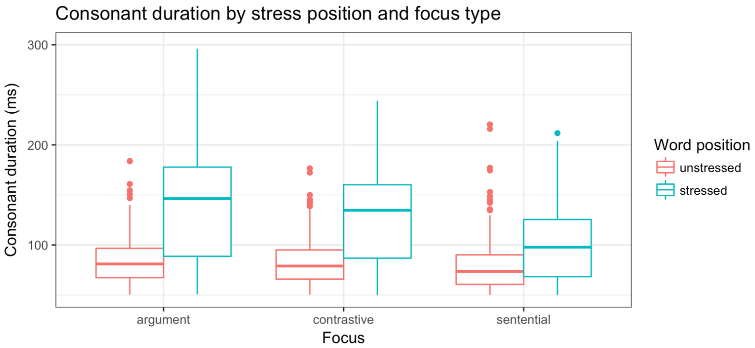
And this lenition is not predictable by rule (the transducer won't help)! Castillo García (2007) notes that there is variable fricative debuccalization, but does not discuss stop voicing/manner lenition.

Stops always have closure in elicited speech (8 speakers, 12 reps per stop) (DiCanio et al, submitted b).



Prosodic structure

Might stress contribute to variable obstruent lenition? Onsets in stressed syllables are longer than unstressed syllables (DiCanio et al, submitted a).



Can we measure voicing automatically?

Obstruent lenition and word position in corpora

While infrequent (5-6% of all cases), certain stops (/k, d/) may be produced as voiced approximants in phrase-final position among AAVE speakers (StoryCorps corpus) (Davidson, 2011).

In Majorcan Spanish, full or partial voicing of voiceless stops /p, t, k/ was observed 35.6% of the time in a spontaneous speech corpus, but 3.7% of the time in a read speech corpus (Hualde et al., 2011). Voicing and lenition of phonologically voiceless stops was not sensitive to word position though.

Subsequent work on Spanish found higher rates of voicing in casual conversational speech (Lewis, 2001; Torreira and Ernestus, 2011).

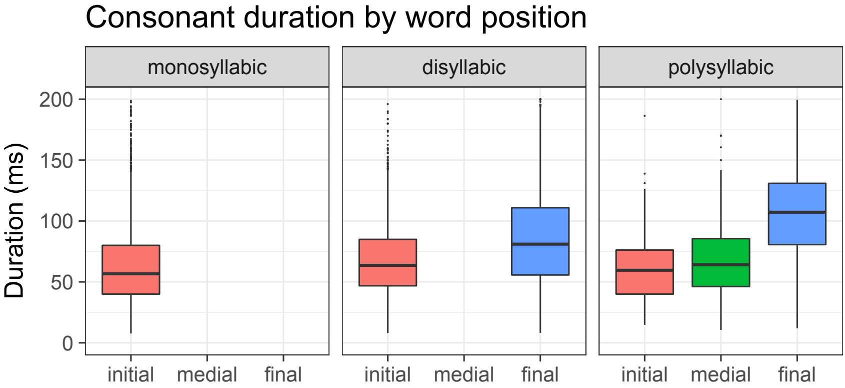
Does the prosodic structure determine the patterns of lenition?

Methods

- Corpus of 6 speakers (3 male, 3 female) producing spontaneous narratives in YM, totalling 107 minutes; force-aligned and corrected.
- Analysis of duration and percentage of voicing during constriction/closure for /t, k, kʷ, s, ʃ, h, tʃ/. Recall that [h] is a free variant of /ʃ, s/.
- A total of 7892 segments were analyzed.
- Hand-labelling of corpus was done in a previous study (DiCanio et al., 2015), but words here were coded by stem position (initial, medial, final syllable), and word size (monosyllabic, disyllabic, polysyllabic).

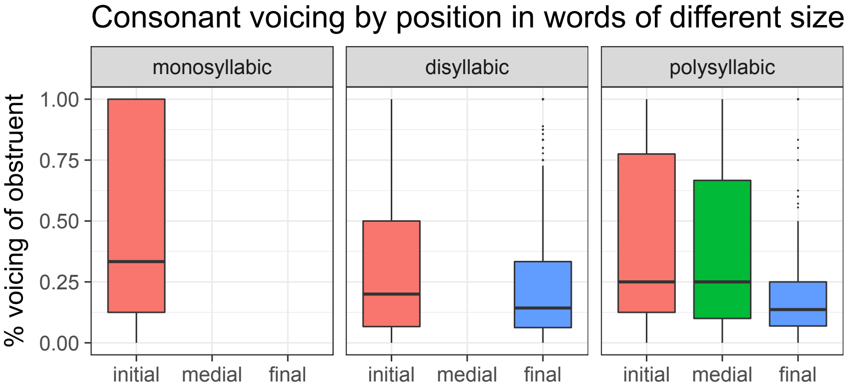
- Duration was extracted with an existing Praat script.
- Voicing was extracted with a script written for Matlab (Chen, W-R). Percentage of voicing during constriction was calculated using a normalized low frequency energy ratio (Kasi and Zahorian, 2002).
- Two separate statistical analyses were run using lmerTest (Kuznetsova et al., 2013), one with duration as the dependent variable and another with percentage of voicing as the dependent variable.
- In each model, word size, word position, and consonant were treated as fixed effects while speaker and item were treated as random effects. No random slopes were included.

Results: duration



A strong effect of position on duration (initial vs. final) was found. Final syllables were longer, more noticeably in polysyllabic words than disyllabic ones.

Results: voicing



Obstruents in word-initial syllables had a larger percentage of voicing than those in word-final syllables.

Discussion

Stem-initial obstruents were both shorter and more likely to be voiced or partially voiced than stem-final obstruents.

Obstruents in stem-final (stressed) syllables were longer than those in word-initial syllables. This matches data from elicited sentences.

Obstruents in stem-final (stressed) syllables were less likely to be partially or fully voiced than those in word-initial syllables.

Unlike languages like English where word-initial position is the locus of domain-related strengthening (Fougeron and Keating, 1997), initial syllables in YM are weakened relative to medial or final syllables.

The prosodic structure of YM partially predicts the degree of voicing observed in the spontaneous speech data.

Predicting surface phonetic variation

Speech style matters for forced alignment and in speech production. Running speech corpora show similar patterns as elicited speech.

Transcription (ELAN) → Phonological transducer → Segmentation by hand for aligner construction → Aligner creation and testing → Word and phone level segmentation

Representational and phonological questions

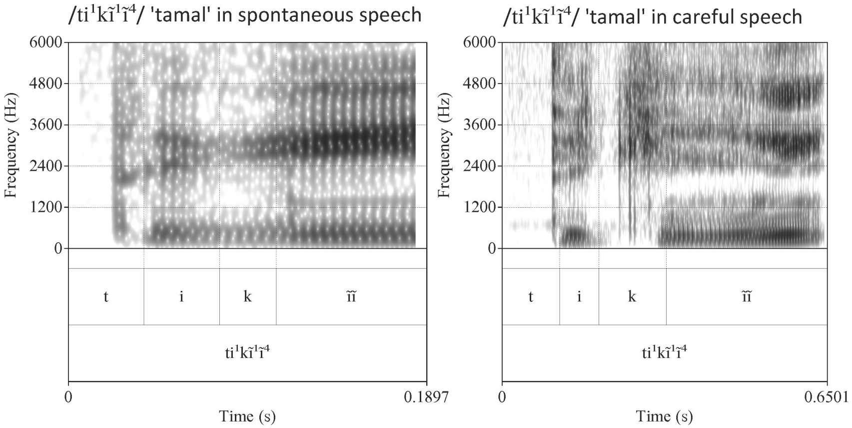
Do we need a language-specific aligner?

You can use an aligner not trained on the language if need be.

Can we predict surface allophonic variation automatically too?

How much phonetic variation occurs?

Obstruents in YM vary in terms of voicing *and* manner.



Qualitative analysis of variation

Examined 89 minutes of corpus used for voicing/lenition study.

Praat script which scanned for the target phone and permitted user to select allophone.

4472 stop tokens (/t, k/) analyzed.

	Vcls stop	Partially vcd stop	Voiced stop	Voiced fric.	Voiced approx.	Nasal	Tap	Deleted
/t/	17.9%	33.0%	21.2%	15.8%	2.7%	6.6%	1.2%	1.6%
/k/	15.3%	20.0%	16.4%	33.5%	7.9%	1.5%	NA	4.8%

Realization	Stop	Voiced
/t/	72.1%	49.1%
/k/	51.7%	64.7%

Variation in manner/voicing for "voiceless unaspirated stops" is very common in running speech. Can a model be trained to detect it? Can we add this as a layer in our speech corpus?

Step	Layer	Method
1	Original transcription	ELAN
2	Surface phonological representation	Phonological transducer
3	Lexical and Phone-level segmentation	Forced alignment
4	**Surface phonetic variation**	???

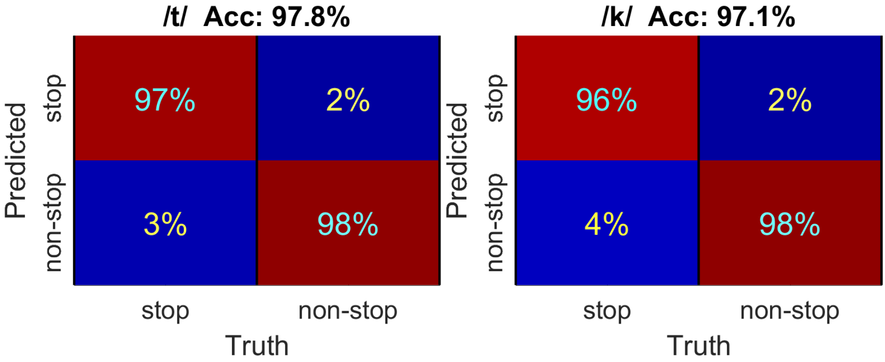
Predicting surface phonetic variation not only permits greater detail in the speech corpus, but allows one to examine low-level variation in speech production without needing to code the acoustic data by hand.

Methods: DNN modelling

- We can use the allophonic labelling from the 4,472 stop tokens to train DNNs (Deep neural networks) to categorize surface phonetic allophones.
- Six models trained: 2-way, 3-way, 4-way models on /t/ and on /k/; (500 nrns) (Hinton et al., 2012).
- 20 MFCC coefficients extracted from each hanning-windowed (10 ms, 2ms step) acoustic signal (48 kHz > 16 kHz) for each stop token. MFCCs were standardized, normalized, and rescaled.
- Models trained on 80% of data, fine-tuned on 10% cross-validation set, and tested on remaining 10% (random split).

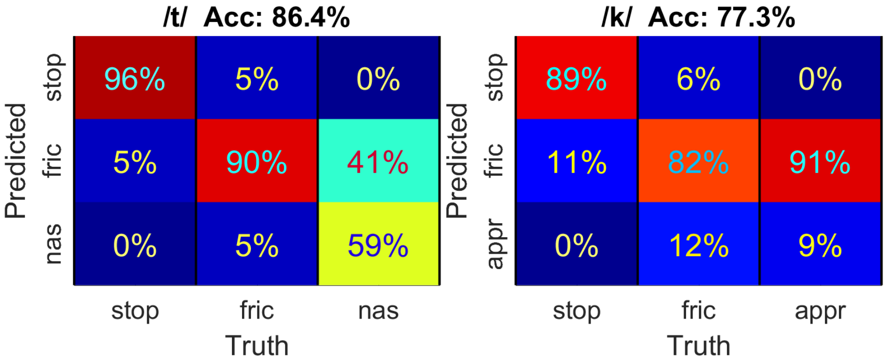
2-way categorization

High accuracy found – stop vs. non-stop



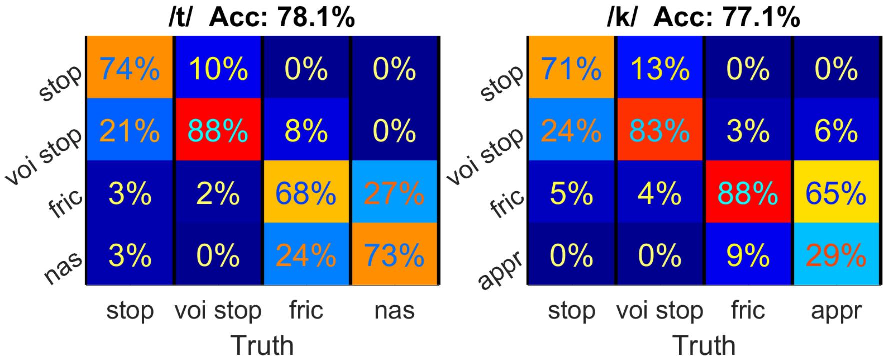
3-way categorization

Higher accuracy found – stop vs. fricative vs. sonorant (nasal or approximant). Sonorant realizations tend to be categorized as fricatives.



4-way categorization

Good accuracy found – voiceless stop vs. voiced stop vs. fricative vs. sonorant (nasal or approximant).



- Despite training on a limited data set, the DNN models showed high accuracy in predicting stop allophones in the test data.
- All models showed excellent stop/continuant identification, though approximants were more poorly identified.
- The four-way model showed good performance in voiceless-voiced stop identification.
- DNN models can detect allophones from continuous speech, which is useful both for improving surface phonetic transcription.

Next steps: compare DNN against simpler models, test on other language data, apply model to corpus data

General discussion

One can adapt an English-based forced aligner to get initial segmentation of a documentation corpus. Speech style matters in speech production and in the choice of the aligner that is used (and in what one trains).

Yet, even after creating a phonological transducer and language-specific aligner, one can observe variation within the surface phonetic representation of the corpus that is not captured.

For YM, prosodic structure explains this variation and it can be modelled based on some relatively simple human categorization data and included as an annotation layer in a speech corpus.

Future directions

- Corpus has been completely transduced and alignment has been checked. ($\sim$ 1 million words)
- Improve DNN performance and expand to other consonant types; include an additional surface phonetic layer.
- Collaborative work at McGill integrating the existing corpus with Speech Corpus Tools.
- Corpus tone production.

ja4bi2 ndio4si2=ni42=un4!

$\text{ku}^{2}\text{ɾu}^{4}\text{a}^{43}\text{=a}^{3}\text{ni}^{2}\text{ʔih}^{5}\text{ɾeʔ}^{1}$!

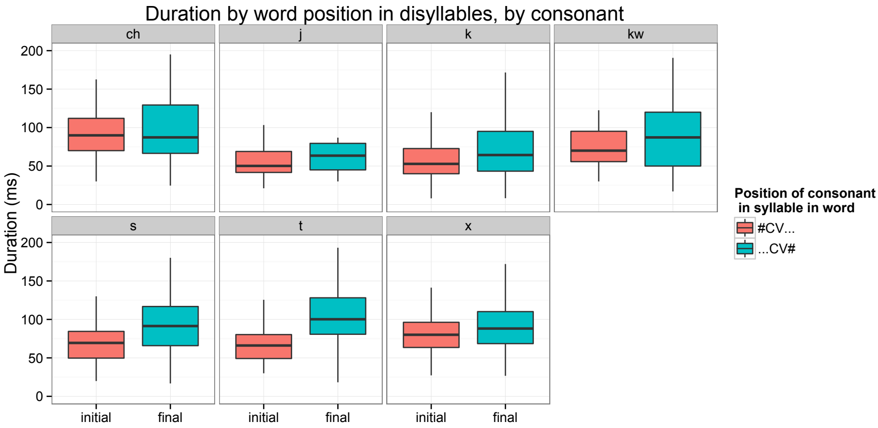
Thank you!

Merci beaucoup!

Gracias!

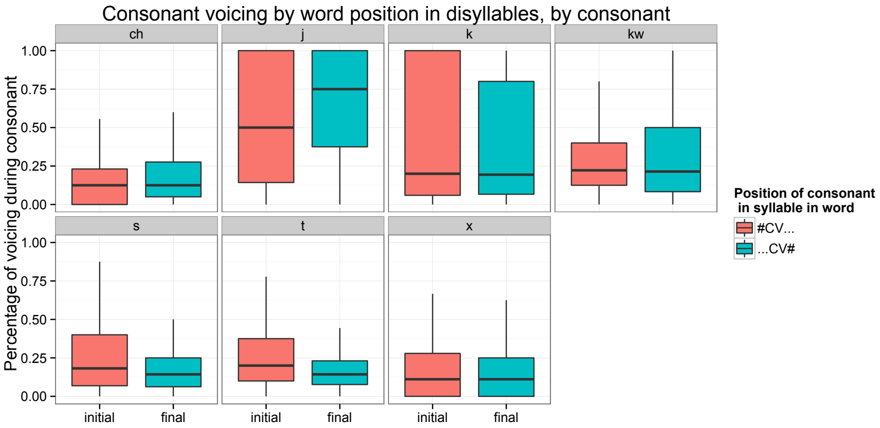
Appendix

Duration effects by consonant in disyllabic words



Appendix

Voicing effects by consonant in disyllabic words



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

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge University Press, 353 pages.

Boersma, P. and Weenink, D. (2016). 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.

Davidson, L. (2011). Characteristics of stop releases in American English spontaneous speech. *Speech Communication*, 53:1042–1058.

DiCanio, C., Amith, J. D., and Castillo García, R. (2014). The phonetics of moraic alignment in Yoloxóchitl Mixtec. In *Proceedings of the* 4^{th} *Tonal Aspects of Language Symposium*. Nijmegen, the Netherlands.

DiCanio, C., Benn, J., and Castillo García, R. (submitteda). The phonetics of information structure in Yoloxóchitl Mixtec.

DiCanio, C., Nam, H., Amith, J. D., Castillo García, R., and Whalen, D. H. (2015). Vowel variability in elicited versus spontaneous speech: evidence from Mixtec. *Journal of Phonetics*, 48:45–59.

DiCanio, C., Nam, H., Whalen, D. H., Bunnell, H. T., Amith, J. D., and Castillo García, R. (2013). Using automatic alignment to analyze endangered language data: Testing the viability of untrained alignment. *Journal of the Acoustical Society of America*, 134(3):2235–2246.

DiCanio, C. and Whalen, D. H. (2015). The interaction of vowel length and speech style in an arapaho speech corpus. In *Proceedings of the 18th International Congress of the Phonetic Sciences*, Glasgow, Scotland.

DiCanio, C., Zhang, C., Whalen, D. H., Amith, J. D., and Castillo García, R. (submittedb). Phonetic structure in Yoloxóchitl Mixtec consonants.

Fougeron, C. and Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America*, 101(6):3728–3740.

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.

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

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., and Sainath, T. N. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97.

Hualde, J. I., Simonet, M., and Nadeu, M. (2011). Consonant lenition and phonological recategorization. *Journal of Laboratory Phonology*, 2:301–329.

Kasi, K. and Zahorian, S. A. (2002). Yet another algorithm for pitch tracking. In *Proceedings of ICASSP02*, pages 361–364. Orlando.

Keating, P. A. and Huffman, M. K. (1984). Vowel variation in Japanese. *Phonetica*, 41:191–207.

Koopmans van Beinum, F. (1980). *Vowel contrast reduction, an acoustic and perceptual study of Dutch vowels in various speech conditions*. PhD thesis, University of Amsterdam, The Netherlands., Academische Pers B. V., Amsterdam.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2013). *lmerTest (R package)*.

Ladefoged, P. (2003). *Phonetic Data Analysis*. Blackwell.

Lewis, A. M. (2001). *Weakening of intervocalic /ptk/ in two Spanish dialects: Toward the quantification of lenition processes*. PhD thesis, University of Illinois at Urbana-Champaign.

Lindblom, B. (1963). Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America*, 35:1773–1781.

Lindblom, B. (1983). The economy of speech gestures. In MacNeilage, P. F., editor, *The Production of Speech*, pages 217–243. Springer-Verlag.

Lindblom, B. (1990). Explaining phonetic variation: a sketch of the H&H theory. In Hardcastle, W. J. and Marchal, A., editors, *Speech Production and Speech Modelling*, pages 403–439. Kluwer Academic Publishers.

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

Meunier, C. and Espesser, R. (2011). Vowel reduction in conversational speech in French: The role of lexical factors. *Journal of Phonetics*, 39:271–278.

Moon, S.-J. and Lindblom, B. (1994). Interaction between duration, context, and speaking style in English stressed vowels. *Journal of the Acoustical Society of America*, 96(1):40–55.

Mücke, D., Nam, H., Hermes, A., and Goldstein, L. (2012). Coupling of tone and constriction gestures in pitch accents. In Hoole, P., Bombien, L., Pouplier, M., Mooshammer, C., and Kühnert, B., editors, *Consonant Clusters and Structural Complexity*, pages 205–230. Walter de Gruyter.

Nam, H., Goldstein, L., Saltzman, E., and Byrd, D. (2004). TADA: An enhanced, portable Task Dynamics model in MATLAB. *Journal of the Acoustical Society of America*, 115(5):2430.

Palancar, E. L., Amith, J. D., and Castillo García, R. (2016). Verbal inflection in Yoloxóchitl Mixtec. In Palancar, E. L. and Léonard, J.-L., editors, *Tone and Inflection: New Facts and New Perspectives*, chapter 12, pages 295–336. Mouton de Gruyter.

R Development Core Team, Vienna, A. (2013). R: A language and environment for statistical computing [computer program], version 3.0.2. http://www.R-project.org, R Foundation for Statistical Computing.

Torreira, F. and Ernestus, M. (2011). Realization of voiceless stops and vowels in conversational French and Spanish. *Journal of Laboratory Phonology*, 2:331–353.

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.