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)

NSF Grant No. 0966411 to Haskins Laboratories
NSF Grant No. 1603323 to the University at Buffalo
The problem: fieldwork → spoken language corpus
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.
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.

DiCanio (UB)  Phon. variation & Mixtec corpus dev.  11/10/17  5 / 55
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)

<table>
<thead>
<tr>
<th></th>
<th>Bilabial</th>
<th>Dental</th>
<th>Alveolar</th>
<th>Post-alveolar</th>
<th>Palatal</th>
<th>Velar</th>
<th>Labialized Velar</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Plosive</strong></td>
<td>p</td>
<td>t</td>
<td></td>
<td></td>
<td></td>
<td>k</td>
<td>k\textsuperscript{w}</td>
</tr>
<tr>
<td><strong>Nasal</strong></td>
<td>m</td>
<td>n</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Post-stopped nasal</strong></td>
<td>m\textsubscript{b}</td>
<td>n\textsuperscript{d}</td>
<td></td>
<td></td>
<td></td>
<td>\eta\textsuperscript{g}</td>
<td></td>
</tr>
<tr>
<td><strong>Tap</strong></td>
<td></td>
<td>r</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Affricate</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>tʃ</td>
<td></td>
</tr>
<tr>
<td><strong>Fricative</strong></td>
<td>β</td>
<td>ñ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Approximant</strong></td>
<td>l</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>j</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Front</th>
<th>Central</th>
<th>Back</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Close</strong></td>
<td>i, ï</td>
<td></td>
<td>u, ŭ</td>
</tr>
<tr>
<td><strong>Close-mid</strong></td>
<td>e, ẽ</td>
<td></td>
<td>o, ō</td>
</tr>
<tr>
<td><strong>Open</strong></td>
<td>a, ă</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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. /ya?4a1/ ‘grey’, /sa?3ma4/ ‘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

<table>
<thead>
<tr>
<th>Morphology</th>
<th>‘to break’ (tr)</th>
<th>‘hang’ (tr)</th>
<th>‘to change’ (intr)</th>
<th>‘to peel’ (tr)</th>
<th>‘to get wet’</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stem</td>
<td>ta3?βi4</td>
<td>tʃi3kū2</td>
<td>na1ma3</td>
<td>kwi1i4</td>
<td>tʃi3i4</td>
</tr>
<tr>
<td>NEG</td>
<td>ta14?βi4</td>
<td>tʃi14kū2</td>
<td>na14ma3</td>
<td>kwi14i14</td>
<td>tʃi14i3</td>
</tr>
<tr>
<td>COMP</td>
<td>ta13?βi4</td>
<td>tʃi13kū2</td>
<td>na13ma3</td>
<td>kwi1i4</td>
<td>tʃi13i3</td>
</tr>
<tr>
<td>INCOMP</td>
<td>ta4?βi4</td>
<td>tʃi4kū2</td>
<td>na4ma13</td>
<td>kwi4i14</td>
<td>tʃi4i4</td>
</tr>
<tr>
<td>1S</td>
<td>ta3?βi42</td>
<td>tʃi3kū2=ju1</td>
<td>na1ma32</td>
<td>kwi1i42</td>
<td>tʃi3i2</td>
</tr>
</tbody>
</table>
Disyllabic words in YM

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

<table>
<thead>
<tr>
<th>Melody</th>
<th>Word</th>
<th>Gloss</th>
<th>Melody</th>
<th>Word</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>ta₁ma₁</td>
<td>without appetite</td>
<td>4.13</td>
<td>na⁴ma₁³</td>
<td>is changing</td>
</tr>
<tr>
<td>1.3</td>
<td>na₁ma₃</td>
<td>to change (intr)</td>
<td>4.14</td>
<td>nda⁴ta₁⁴</td>
<td>is splitting up</td>
</tr>
<tr>
<td>1.4</td>
<td>na₁ma₄</td>
<td>soap</td>
<td>4.24</td>
<td>ya⁴ma²⁴</td>
<td>Amuzgo person</td>
</tr>
<tr>
<td>1.32</td>
<td>na₁ma₃²</td>
<td>I will change myself</td>
<td>4.42</td>
<td>na⁴ma⁴²</td>
<td>I often pile rocks</td>
</tr>
<tr>
<td>1.42</td>
<td>na¹ma⁴²</td>
<td>my soap</td>
<td>13.2</td>
<td>hi¹³ni²</td>
<td>has seen</td>
</tr>
<tr>
<td>3.2</td>
<td>na³ma²</td>
<td>wall</td>
<td>13.3</td>
<td>na¹³na³</td>
<td>has photographed oneself</td>
</tr>
<tr>
<td>3.3</td>
<td>na³ma³</td>
<td>to change (tr)</td>
<td>13.4</td>
<td>na¹³ma⁴</td>
<td>has piled rocks</td>
</tr>
<tr>
<td>3.4</td>
<td>na³ma⁴</td>
<td>sprout</td>
<td>14.2</td>
<td>na¹⁴ma²</td>
<td>I will not change</td>
</tr>
<tr>
<td>3.42</td>
<td>na³ma⁴²</td>
<td>I will pile rocks</td>
<td>14.3</td>
<td>na¹⁴ma³</td>
<td>to not change</td>
</tr>
<tr>
<td>4.1</td>
<td>ka⁴nda¹</td>
<td>is moving (intr)</td>
<td>14.4</td>
<td>na¹⁴ma⁴</td>
<td>to not pile rocks</td>
</tr>
<tr>
<td>4.2</td>
<td>na⁴ma²</td>
<td>I am changing</td>
<td>14.13</td>
<td>na¹⁴ma¹³</td>
<td>to not change oneself</td>
</tr>
<tr>
<td>4.3</td>
<td>na⁴ma³</td>
<td>it is changing</td>
<td>14.14</td>
<td>nda¹⁴ta¹⁴</td>
<td>to not split up</td>
</tr>
<tr>
<td>4.4</td>
<td>na⁴ma⁴</td>
<td>is piling rocks</td>
<td>14.42</td>
<td>na¹⁴ma⁴²</td>
<td>I will not pile rocks</td>
</tr>
</tbody>
</table>

(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.

<table>
<thead>
<tr>
<th>Mixtec</th>
<th>P2FA</th>
<th>hmAlign</th>
</tr>
</thead>
<tbody>
<tr>
<td>/t/ [t]</td>
<td>T $[t^h, t, \tilde{t}, r]$</td>
<td>TT $[t]$</td>
</tr>
<tr>
<td>/k/ [k]</td>
<td>K $[k^h, k]$</td>
<td>KK $[k]$</td>
</tr>
<tr>
<td>/kʷ/ [kʷ]</td>
<td>K $[k^h, k]$</td>
<td>KK $[k]$</td>
</tr>
<tr>
<td>/ʔ/ [ʔ]</td>
<td>T $[t^h, t, \tilde{t}, r]$</td>
<td>TQ $[\tilde{t}]$</td>
</tr>
</tbody>
</table>
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. /ko₁o₄/ ‘snake’, /nᵈa¹βa¹/ ‘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
Results: general

Agreement is better with hm-Align than with P2FA.

<table>
<thead>
<tr>
<th>Threshold</th>
<th>P2FA</th>
<th>hm-Align</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 ms</td>
<td>32.3%</td>
<td>40.6%</td>
</tr>
<tr>
<td>20 ms</td>
<td>52.3%</td>
<td>61.4%</td>
</tr>
<tr>
<td>30 ms</td>
<td>65.7%</td>
<td>70.9%</td>
</tr>
<tr>
<td>40 ms</td>
<td>74.8%</td>
<td>81.2%</td>
</tr>
<tr>
<td>50 ms</td>
<td>79.6%</td>
<td>86.7%</td>
</tr>
</tbody>
</table>

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

![Agreement by consonant class at end points](image)

DiCanio (UB) Phon. variation & Mixtec corpus dev.

11/10/17 17 / 55
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.

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Elicited Speech</th>
<th>Running Speech</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>hm-Align</td>
<td>P2FA</td>
</tr>
<tr>
<td>10 ms.</td>
<td>40.6%</td>
<td>32.3%</td>
</tr>
<tr>
<td>20 ms.</td>
<td>61.4%</td>
<td>52.3%</td>
</tr>
<tr>
<td>30 ms.</td>
<td>70.9%</td>
<td>65.7%</td>
</tr>
<tr>
<td>40 ms.</td>
<td>81.2%</td>
<td>74.8%</td>
</tr>
<tr>
<td>50 ms.</td>
<td>86.7%</td>
<td>79.6%</td>
</tr>
</tbody>
</table>
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 (preceeding 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

DiCanio (UB)  Phon. variation & Mixtec corpus dev.  11/10/17  27 / 55
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 $\text{ki' in}^3 = \text{on}^4$ /$\text{kǐ?i}^3 = \text{ṅ}^4$/ ‘you take’ is pronounced $[\text{kjn?i}^3]$?

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. $\text{ki' in}^3 = \text{on}^4$ Input
2. $\text{ki' on}^4$ Vowel replacement/harmony
3. $\text{ki' on}^4$ Replace all preceding vowels if [-high]
   else [+high] $\rightarrow$ glide
4. $\text{kyo' on}^4$ 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.’ \([ti^1\tilde{y}^1\tilde{a}^4 \text{ ja}^4 \text{ du}^3\text{ yu}^3 \text{ r}^3\tilde{a}^4]\) (left) vs. \([ti^1\tilde{k}^1\tilde{i}^4 \text{ ja}^3 \text{ tu}^3\text{ ku}^3 \text{ r}^3\tilde{a}^4]\) (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, kʰ, 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.
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

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.

/ti₁kᵢᵢ⁴/ 'tamal' in spontaneous speech

/ti₁kᵢᵢ⁴/ 'tamal' in careful speech
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.

<table>
<thead>
<tr>
<th></th>
<th>Vcls stop</th>
<th>Partially vcd stop</th>
<th>Voiced stop</th>
<th>Voiced fric.</th>
<th>Voiced approx.</th>
<th>Nasal</th>
<th>Tap</th>
<th>Deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td>/t</td>
<td>17.9%</td>
<td>33.0%</td>
<td>21.2%</td>
<td>15.8%</td>
<td>2.7%</td>
<td>6.6%</td>
<td>1.2%</td>
<td>1.6%</td>
</tr>
<tr>
<td>/k</td>
<td>15.3%</td>
<td>20.0%</td>
<td>16.4%</td>
<td>33.5%</td>
<td>7.9%</td>
<td>1.5%</td>
<td>NA</td>
<td>4.8%</td>
</tr>
</tbody>
</table>

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?

<table>
<thead>
<tr>
<th>Step</th>
<th>Layer</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Original transcription</td>
<td>ELAN</td>
</tr>
<tr>
<td>2</td>
<td>Surface phonological representation</td>
<td>Phonological transducer</td>
</tr>
<tr>
<td>3</td>
<td>Lexical and Phone-level segmentation</td>
<td>Forced alignment</td>
</tr>
<tr>
<td>4</td>
<td><strong>Surface phonetic variation</strong></td>
<td>???</td>
</tr>
</tbody>
</table>

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

/t/ Acc: 97.8%

Predicted

<table>
<thead>
<tr>
<th></th>
<th>stop</th>
<th>non-stop</th>
</tr>
</thead>
<tbody>
<tr>
<td>stop</td>
<td>97%</td>
<td>3%</td>
</tr>
<tr>
<td>non-stop</td>
<td>2%</td>
<td>98%</td>
</tr>
</tbody>
</table>

/k/ Acc: 97.1%

Predicted

<table>
<thead>
<tr>
<th></th>
<th>stop</th>
<th>non-stop</th>
</tr>
</thead>
<tbody>
<tr>
<td>stop</td>
<td>96%</td>
<td>4%</td>
</tr>
<tr>
<td>non-stop</td>
<td>2%</td>
<td>98%</td>
</tr>
</tbody>
</table>
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).

![Accuracy table for /t/ and /k/ sounds]

- /t/ Acc: 78.1%
  - stop: 74%
  - voiceless stop: 21%
  - fric: 3%
  - nas: 3%

- /k/ Acc: 77.1%
  - stop: 71%
  - voiceless stop: 24%
  - fric: 5%
  - nas: 3%
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. (∼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!
ku^2ru^4a^43=a^3ni^2?ih^5re^1!
Thank you!
Merci beaucoup!
Gracias!
Appendix

Duration effects by consonant in disyllabic words

DiCanio (UB)  Phon. variation & Mixtec corpus dev.  11/10/17  54 / 55
Appendix

Voicing effects by consonant in disyllabic words

Consonant voicing by word position in disyllables, by consonant

Position of consonant in syllable in word
- #CV...
- ...CV#

Percentage of voicing during consonant

Initial vs. final

DiCanio (UB)  Phon. variation & Mixtec corpus dev.  11/10/17  55 / 55


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


