Uneven success: Automatic speech recognition and ethnicity-related dialects

Alicia Beckford Wassink
Department of Linguistics, University of Washington
https://depts.washington.edu/sociolab/

Panel Title:
Ethical risks of voice technology: A sociolinguistic perspective
14 February 2020
Outline

I. Research Aims

II. Background
   a) What’s “sociophonetics?”
   b) Our tool: CLOx

III. Methods
   a) The sample: 4 ethnic groups from Pacific Northwest English (PNWE) study corpus
   b) Targeted linguistic variables

IV. By-ethnicity results

V. Some surprising findings

VI. Conclusions
acknowledgements

CLOx Team:
- Campion Fellin
- David Nichols
- Robert Squizzero

PNWE Team:
- Isabel Bartholomew
- Sophia Chan
- Cady Gansen
- Monica Jensen
- Nathan Johnson
- Michael Scanlon

National Science Foundation
BCS-1844350
I. Aims

• Not all features of speech are handled well
• Contemporary use cases:
  • Siri, Alexa, Cortana
  • Payment-by-phone
• Inequity in access to services

• Research Questions: What differences do we observe in error types? What dialect features appear to be most challenging for our CLOx system?
II. Background
Sociophonetics

• A subfield of linguistics that identifies and explains socially-structured variation in the sound systems of human languages.

• Concerned with how such variation is learned, stored cognitively, subjectively evaluated, and processed in speaking and listening.

Foulkes, Scobbie and Watt 2010; diPaolo and Yaeger-Dror 2011
Linguistic variable

• Def.: “a linguistic form whose occurrence cannot be explained without taking social characteristics into account”

• Ex. “They were the tawatees.”

Lexical variable
“local doctor, medicine person”

[deɪ wə dɪ tawətɪz]
Phonetic variable
International Phonetic Alphabet (IPA)
(th)-stopping

Yakama English (WA)
Reading Passage example

Vowels:
A formant is a concentration of acoustic energy around a particular frequency in the speech signal.

African American (F)

<table>
<thead>
<tr>
<th></th>
<th>F3</th>
<th>F2</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>w</td>
<td>u</td>
<td>?</td>
</tr>
<tr>
<td>b</td>
<td>i</td>
<td>h</td>
<td>æ</td>
</tr>
<tr>
<td>p</td>
<td>ij</td>
<td>æ</td>
<td>?</td>
</tr>
<tr>
<td>t</td>
<td>h</td>
<td>aw</td>
<td>s</td>
</tr>
</tbody>
</table>

I WOULD BE HAPPY IN THAT HOUSE

Yakama (M)

<table>
<thead>
<tr>
<th></th>
<th>aɪ</th>
<th>w</th>
<th>o</th>
<th>d</th>
<th>b</th>
<th>i</th>
<th>h</th>
<th>æ</th>
<th>p</th>
<th>ij</th>
<th>ɪ</th>
<th>nd</th>
<th>æ</th>
<th>?</th>
<th>h</th>
<th>œ</th>
<th>m</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>WOULD</td>
<td>BE</td>
<td>HAPPY</td>
<td>IN</td>
<td>THAT</td>
<td>HOME</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
III. Methods
Speaker sample: 4 WA dialects

Map credit: nationalatlas.gov ©2019: US Geographical Survey
Note: Speaker classification into ethnic groups was based upon:

• Speaker’s self-identification
• Social network data (membership in a speech community)
• Length of time in speech community

* Neither dialect nor ethnic affiliation may be definitively ascertained by visual appearance.
Tasks

• 16 speakers, 4 Ethnic groups
• Three tasks:
  • Dyadic conversations (casual, most dialectal forms)
  • Reading Passage (read, common forms)
  • Word Game task (unscripted, common forms)
• Data amounts:
  • Approx. 45 - 90 min. of speech per recording
  • 6,654 - 16,276 words per ethnic group
• Submitted to ASR tool
• Coding:
  • Manual coding in Praat (acoustic analysis software). Auditory analysis supplemented by use of waveform and spectrogram
Our Tool: CLOx

- **Client Libraries Oxford**
- Automated audio transcription service for linguists developed by the Sociolinguistics Laboratory at the University of Washington.
- Built on the Microsoft Speech Service (via Azure subscription to Cognitive Services).
- Automatic speech recognition uses the Speech-to-text service SDK.
- CLOx delivers a conversational recording to MS Speech, which returns plain-text transcribed output, then CLOx performs output checking and supplies timestamps indicating the start and end time of each run of speech.
- We estimate that CLOx transcription is at least **five times faster** than manual transcription (hence, the logo!)
Our Tool: CLOx
# General error types

<table>
<thead>
<tr>
<th>Code</th>
<th>Label</th>
<th>Example error</th>
<th>Target</th>
<th>IPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>reduction</td>
<td>lotta</td>
<td>lot of</td>
<td>varies</td>
</tr>
<tr>
<td>D</td>
<td>disfluencies</td>
<td>enough</td>
<td>and uh</td>
<td></td>
</tr>
<tr>
<td>NC</td>
<td>no code</td>
<td>changing</td>
<td>digging</td>
<td>ø</td>
</tr>
<tr>
<td>NULL</td>
<td>words inserted</td>
<td>could (&quot;windows could they would&quot;)</td>
<td>topless</td>
<td></td>
</tr>
<tr>
<td>PN</td>
<td>Proper name</td>
<td>topless</td>
<td>Toppenish</td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>Homophone</td>
<td>are<del>R</del>our</td>
<td>are<del>R</del>our</td>
<td></td>
</tr>
</tbody>
</table>

- Not associated with any specific dialect
- Not targeted for sociophonetic study
# Sociolinguistic Variables

## Consonants:

<table>
<thead>
<tr>
<th>Code</th>
<th>Sociolinguistic Label</th>
<th>Example error</th>
<th>Target</th>
<th>IPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>(ing)</td>
<td>-ing (unstressed)</td>
<td>pick into</td>
<td>picking too</td>
<td>[ɪŋ] vs [ɪn] vs [ɪn]</td>
</tr>
<tr>
<td>(TH)</td>
<td>th-stopping</td>
<td>den</td>
<td>then</td>
<td>/ð/ → [d]</td>
</tr>
<tr>
<td>(?)</td>
<td>word-medial glottalization</td>
<td>right are</td>
<td>writer</td>
<td>/t/ → [ʔ]</td>
</tr>
<tr>
<td>(u)</td>
<td>coda-r deletion</td>
<td>what a</td>
<td>water</td>
<td>/ɾ/ → ∅</td>
</tr>
<tr>
<td>(d)</td>
<td>consonant cluster deletion</td>
<td>pace [peɪs]</td>
<td>paced /peɪst/</td>
<td>/st/ → [s]</td>
</tr>
<tr>
<td>(l)</td>
<td>lenition</td>
<td>sheep</td>
<td>cheap</td>
<td>/tʃ/ → [ʃ]</td>
</tr>
</tbody>
</table>

- **ARE** associated with specific dialects
- **ARE** targeted for sociophonetic study
## Sociolinguistic Variables

### Vowels:

<table>
<thead>
<tr>
<th>Code</th>
<th>Sociolinguistic Label</th>
<th>Example error</th>
<th>Target</th>
<th>IPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>(ɪ)</td>
<td>(ɪ)-tensing</td>
<td>peaking</td>
<td>picking</td>
<td>/ɪ/ → [i]</td>
</tr>
<tr>
<td>(ɔ)</td>
<td>caught/cot merger</td>
<td>com, cot</td>
<td>calm, caught</td>
<td>/ɔ/ → [a], /ɔ/ → [a]</td>
</tr>
<tr>
<td>(æg)</td>
<td>pre-voiced velar (æ)-raising</td>
<td>beg</td>
<td>bag</td>
<td>/æg/ → [e:g]</td>
</tr>
<tr>
<td>(æ)</td>
<td>mistaking (æ) for other Vowel</td>
<td>infect</td>
<td>in fact</td>
<td>/æ/ → [a], /æ/ → [ɛ]</td>
</tr>
<tr>
<td>(ɛg)</td>
<td>pre-voiced velar (ɛ)-raising</td>
<td>beg</td>
<td>bake</td>
<td>/ɛg/ → [e:g]</td>
</tr>
<tr>
<td>(ɻ)</td>
<td>(ɻ)-raising</td>
<td>is</td>
<td>us</td>
<td>/ɻ/ → [i], /ɻ/ → [ɻ]</td>
</tr>
<tr>
<td>(ow)</td>
<td>(ow)-fronting</td>
<td>boot</td>
<td>boat</td>
<td>/ow/ → [u]</td>
</tr>
<tr>
<td>(prel)</td>
<td>prelateral back vowel merger</td>
<td>full, hole</td>
<td>fool, hull</td>
<td>/ul/ ↔ /ol/, /ul/ ↔ /ul/, /ul/ ↔ /ol/</td>
</tr>
<tr>
<td>(IN)</td>
<td>pin/pen merger</td>
<td>pin</td>
<td>pen</td>
<td>/IN/ ↔ /ɛn/</td>
</tr>
<tr>
<td>V</td>
<td>other vowel error</td>
<td>greet</td>
<td>great</td>
<td>varies</td>
</tr>
<tr>
<td>O</td>
<td>other (phonetic/phonological errors) thing, faults</td>
<td>vague, false</td>
<td>varies</td>
<td></td>
</tr>
</tbody>
</table>

- **ARE** associated with specific dialects
- **ARE** targeted for sociophonetic study
CLOx Errors, by type (Caucasian American Subsample)

General Errors
- H
- PN
- NULL
- NC
- D
- R
- O

Vowel Errors
- (IN)
- (prel)
- (ow)
- (æ)
- (æg)
- (i)
- (e)
- (r)
- (l)
- (d)
- (j)
- (ʔ)
- (ING)

Consonant Errors
- Watch this space!

Legend:
- % 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 0% 9% 36% 16% 4% 1% 4% 14% 16%
Normalized Frequency ($nf$)

- $E$: Erroneous forms across all targeted linguistic variables in a corpus
- $N$: Total word count for the corpus
- $B$: Base of normalization = 100 words
- $nf$: Number of error in corpus / total corpus x base of normalization

$$nf = \frac{E}{N} \times B$$

$$E = 668$$
$$N = 16,276$$
$$nf = \frac{668}{16,276} \times 100$$
$$= 4.104$$
## IV. Results

- Overall $nf$, by ethnicity

<table>
<thead>
<tr>
<th>Group</th>
<th>N=</th>
<th>$nf$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caucasian American</td>
<td>6,654</td>
<td>1.5</td>
</tr>
<tr>
<td>African American</td>
<td>16,276</td>
<td>4.1</td>
</tr>
<tr>
<td>Chicanx</td>
<td>3,986</td>
<td>8.8</td>
</tr>
<tr>
<td>Yakama</td>
<td>14,581</td>
<td>8.9</td>
</tr>
</tbody>
</table>
#1: Fewest errors \((nf=1.5)\)

CLOx Errors, by type (Caucasian American Subsample)

<table>
<thead>
<tr>
<th>Type</th>
<th>Target</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>&quot;where’s&quot;</td>
<td>&quot;worse&quot;</td>
</tr>
<tr>
<td>O</td>
<td>“grading”</td>
<td>“grating”</td>
</tr>
<tr>
<td>V</td>
<td>“well”</td>
<td>“will”</td>
</tr>
<tr>
<td>(æg)</td>
<td>“bag”</td>
<td>“beg”</td>
</tr>
</tbody>
</table>

Target

<table>
<thead>
<tr>
<th>H</th>
<th>PN</th>
<th>NULL</th>
<th>NC</th>
<th>D</th>
<th>R</th>
<th>O</th>
<th>V</th>
<th>(IN)</th>
<th>(prel)</th>
<th>(ow)</th>
<th>(ʔ)</th>
<th>(æ)</th>
<th>(æg)</th>
<th>(o)</th>
<th>(ɪ)</th>
<th>(l)</th>
<th>(d)</th>
<th>(u)</th>
<th>(?)</th>
<th>(TH)</th>
<th>(ing)</th>
</tr>
</thead>
<tbody>
<tr>
<td>%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>1%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
</tbody>
</table>
CLOx Errors, by type (African American Subsample)

9 additional error types, including:
(V): “head” → “had”
(IN): “pen” → “pin”
(prel): “Tyler Powell” → “taller dollar”
(ʔ): “dawn” → “done”
CLOx Errors, by type (Chicanx Subsample)

(!): “sheet” → “cheat”
(?): “a kitten” → “akin”
(εg): “peg” → “pig”
#4: Most errors ($nf=8.9$)

CLOx Errors, by type (Yakama Subsample)

(TH): “pen them” ➞ “pendam”
(ing): “diggin” ➞ “dig in”
Some surprises

<table>
<thead>
<tr>
<th>Target</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Northwest</td>
<td>Earth less</td>
</tr>
<tr>
<td>Northwesterner</td>
<td>Northwestern Scenario Northwest Eric</td>
</tr>
<tr>
<td>Me</td>
<td>Maine</td>
</tr>
<tr>
<td>Certain [sɨʔɪn]</td>
<td><em>no error</em></td>
</tr>
<tr>
<td>hooman</td>
<td>Whom</td>
</tr>
<tr>
<td>Jobs for</td>
<td>Javascript</td>
</tr>
<tr>
<td>A lot of it</td>
<td>online</td>
</tr>
</tbody>
</table>
Conclusions

• This research has accomplished a cross-ethnicity comparison of dialect-based ASR performance
  • Important! Quantified contribution of linguistic variables to error profile
• Is leveraging sociolinguistic knowledge of the fine phonetic detail in dialect variation worth it? Yes!
  • Eliminate approximately 26% of observed errors
• Worthwhile for linguists, too. ASR is a useful tool on the way to “actual” linguistic analysis.
• Not fast (sociophonetic analysis automated for vowels, not for consonants, not for non-majority dialects)

• Room for collaboration on transcription error reduction
• Room to improve access for people to services that rely increasingly upon ASR.
<table>
<thead>
<tr>
<th>Rank</th>
<th>Error</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.</td>
<td>pza</td>
<td>pa</td>
</tr>
<tr>
<td>9.</td>
<td>izic</td>
<td>Isaac</td>
</tr>
<tr>
<td>8.</td>
<td>arndern</td>
<td>and during</td>
</tr>
<tr>
<td>7.</td>
<td>woon did</td>
<td>wounded</td>
</tr>
<tr>
<td>6.</td>
<td>Freycinet</td>
<td>A feast isn’t it?</td>
</tr>
<tr>
<td>5.</td>
<td>anfang</td>
<td>fawn</td>
</tr>
<tr>
<td>4.</td>
<td>edgecator</td>
<td>educator</td>
</tr>
<tr>
<td>3.</td>
<td>plagge</td>
<td>plague</td>
</tr>
<tr>
<td>2.</td>
<td>Lenny Edge</td>
<td>lineage</td>
</tr>
<tr>
<td>1.</td>
<td>Grandpa Minecraft</td>
<td>Grandpa minded</td>
</tr>
</tbody>
</table>

Error Target
Thank you!

wassink@uw.edu
Slides: https://depts.washington.edu/sociolab

Try CLOx:
https://clox.ling.washington.edu/
References
