Uneven success: Racial Bias in Automatic Speech Recognition

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

University of Michigan
18 January 2021
Rev. Dr. Martin Luther King, jr. Colloquium
Outline

Acknowledgements

Aims of the Talk

Background
  What do I mean by racial bias?
  Where do we see bias in language-related systems?

Methods
  Our tool: CLOx
  The sample: 4 ethnic groups from Pacific Northwest English (PNWE) study corpus
  Targeted linguistic variables

By-ethnicity results

Some surprising findings

Conclusions
acknowledgements

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

PNWE Team:
- Isabel Bartholomew
- Sophia Chan
- Cady Gansen

Not pictured:
- Jake McManus
- Amina Venton
- Diana Davidson

Monica Jensen
Nathan Johnson
Michael Scanlon

National Science Foundation
BCS-1844350
Equality Now: the president has the power

The new administration has the opportunity to be the first in 100 years of American history to adopt a radically new approach to the question of civil rights. It must begin, however, with the firm conviction that the principle is no longer in doubt. The day is past for tolerating vicious and inhuman opposition on a subject which determines the lives of twenty million Americans. We must decide that in a new era, there must be a new thinking. If we fail to make this positive decision, an awakening world will conclude that we have become a fossil nation, morally and politically; and no floods of refrigerators, automobiles or color television sets will rejuvenate our image.

Aims of this project

• Support for the larger PNWE research study
• Not all features of speech are handled well
• Contemporary use cases:
  • Siri, Alexa, Cortana
  • Payment-by-phone, OnStar
• Inequity in access to services
• Knowledge regarding sociolinguistic variation has yet to be exploited in acoustic model architectures
• Personal and professional significance for me: an area in which to pursue equity

Research Questions:

1. Is there a difference in error rates for four ethnicity-related subsamples?  
   If so, what differences do we observe in error rate?  
   What is the by-ethnicity distribution of phonetic error types?

2. What dialect features appear to be most challenging for our CLOx speech-to-text service (Microsoft)?  
   Are these dialect features more typically found in the more casual speech tasks?
Background
What do I mean by racial bias?

• A form of implicit bias
  • Automatic associations or stereotypes made by individuals in the unconscious state of mind.
  • No explicit intent to harm
  • Associations influence behavior, “making people respond in biased ways even when they are not explicitly prejudiced.”

   National Initiative for Building Community Trust and Justice (2015)

• Defined for organizations
  • 1) Unequal access to the beneficial work of the organization, 2) Racial disparities in the structure of the organization in roles and offices, 3) Systematic pattern of inclusion and exclusion, or hierarchical distinction, in how the work proceeds, 4) Failure to examine disparities with intent to identify, address or reverse underlying causes

   Maryfield (2018), Justice Research and Statistics Association
   Charity Hudley (2017)
Racial bias in Linguistics?

- Language as part of the “master narrative” of cultural description
  - Linguistic categories were used to elaborate a set of cultural categories for humankind
  - Focus on languages as if these were monolithic
  (Hutton, 1999)

- Classification of language groups centering a monolingual ideal
  - even sociolinguists!
  - NORMs: non-mobile, older, rural, (majority ethnicity) males

- Beliefs about who is and is not a “typical” member of a language group or speech community based upon analysts’ assessment of speaker race
Colonial bias in Linguistics?

- Examining Native American language varieties only through an “endangerment lens”
  - What constitutes a native speaker?
  - What constitutes “knowing” a language?
  - Decolonized approaches to addressing language shift and language *return*  
    (Leonard, 2019)

- Exclusion of other varieties spoken in Native American communities  
  (American English sociolects)

- For the PNWE study, inclusion of Yakama English allows:  
  - Departure from dictum to hold certain speakers aside until after that primary work is done  
  - Sophisticated study of sociolectal features (transfer from heritage language)  
  - Participation in regional Pacific Northwest forms
Racial bias in Language-related technology?

Koenecke, et al. (2020)

- Contemporaneous with the PNWE ASR study, Stanford study of Word Error Rates (WERs) in sociolinguistic corpora of AAE speech
  - 5 ASR systems (Google, Amazon, Apple, IBM, Microsoft)
  - only previous sociolinguistic study of racial bias in ASR system performance
  - Syntactic constructions (copula deletion “He a pastor.”)

- Examination of **perplexity**:
  - **Def.**: In language models, the number of reasonable continuations of a phrase
  - Language model **not** prone to bias (perplexity *lower* for AAE than GAE), even though high WERs were observed.

- **Results “must be due” to phonetic factors**

  Ex. “the dog jumped over the_____.”

<table>
<thead>
<tr>
<th>Fence</th>
<th>Box</th>
<th>Stick</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perplexity=3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Speech Recognition: primer

• Black box problem, but architecture is probably something like ...

Speech preprocessing → MFCC features → Acoustic model → Pronunciation model → Language model

X = x₁ x₂...xₜ

Grande:/ɡɹɑndeɪ/
Hazelnut: /heɪzəlnət/
Latte: /lʌtθɛɪ/

“grande hazelnut latte”

Leverage sociophonetic knowledge?

Adapted from Hui (2019)
Methods

Talkers
16 speakers, 4 Ethnic groups
- Yakima (4 M, 2 F)
- Mexican American (2 M, 1 F)
- African American (1 M, 2 F)
- Caucasian American (1 M, 3 F)

Data amounts
- Approx. 45 - 90 min. of speech per recording
- Minimum of 20 min. of speech per talker
- 9,174 - 22,773 words per ethnic group

Corpus
- 13 hours (4.99 GB)

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
Speaker sample: 4 WA dialects

Map credit: nationalatlas.gov ©2019: US Geographical Survey
# Tasks

## Three tasks:

<table>
<thead>
<tr>
<th>Task</th>
<th>Style</th>
<th>Common Lexical content?</th>
<th>Task Word Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Free-flowing speech</td>
<td>Casual (dyadic)</td>
<td>Uncontrolled (common topics, QGenII)</td>
<td>517-6019</td>
</tr>
<tr>
<td>Lexical Task*</td>
<td>Semi-casual (individual)</td>
<td>Semi-Controlled</td>
<td>218-691</td>
</tr>
<tr>
<td>Reading passage “The Cat and the Mice” (Aesop’s Fables)</td>
<td>Citation</td>
<td>Controlled</td>
<td>342 (fixed)</td>
</tr>
</tbody>
</table>

**Lexical task (word games):**
- Lists (numbers, days of the week, breakfast foods, farm animals)
- Minimal pairs (dawn/don)
- Semantic differentials (what is the difference in meaning between a “sack” and a “bag”?)
Our Tool: CLOx

- **Client Libraries Oxford**

- Automated audio transcription service for linguists developed by the Sociolinguistics Laboratory at the University of Washington.

- Automatic speech recognition uses the Speech-to-text service SDK (Microsoft Cognitive Services, Speech Division).

- CLOx delivers a conversational recording to MS, 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.

https://clox.ling.washington.edu/
Our Tool: CLOx
Data Handling

• All recordings submitted to ASR tool (CLOx)
• Transcripts returned by CLOx were manually coded for errors
  • Each recording was audited using ELAN, errors manually entered into an Excel database
  • Erroneous phone
  • Intended phone
• Inter-rater reliability (agreement in coding over 20% of each file)
Phonetic Error Rate (PER)

Normalized frequency measure, calculated as the proportion of all errors falling into a particular sociolinguistic variable class.

\[ E \quad \text{Erroneous forms across all targeted linguistic variables in a corpus} \]

\[ N \quad \text{Total word count for the corpus} \]

\[ B \quad \text{Base of normalization} = 100 \text{ words} \]

\[ nf = (E/N) \times B \]

Number of error in corpus / total corpus x base of normalization

\[ E = 668 \]

\[ N = 16,276 \]

\[ nf = (668/16276) \times 100 \]

\[ = 4.104 \]
## 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></td>
</tr>
<tr>
<td>NULL</td>
<td>words inserted</td>
<td>could (&quot;windows could they would&quot;)</td>
<td>Ø</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
Targeted 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 [ɪŋ] 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>/ɹ/ → ϕ</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>/tf/ → [ʃ]</td>
</tr>
</tbody>
</table>

Why a common set of variables?

• Assess extent to which regional changes present a problem for ASR
• We know that some forms span non-standard dialects of English
• It may be that certain errors are particular to certain sociolects
• If we see common errors for multiple groups, inclusion in the AM will represent greater gains for ASR.

Wassink (2017), Wassink and Hargus (2020)
Sociolinguistic Variables

Vowels:

<table>
<thead>
<tr>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>(ɪ)</td>
</tr>
<tr>
<td>(ʊ)</td>
</tr>
<tr>
<td>(æ)</td>
</tr>
<tr>
<td>(ɛ)</td>
</tr>
<tr>
<td>(ʌ)</td>
</tr>
<tr>
<td>(ow)</td>
</tr>
<tr>
<td>(prel)</td>
</tr>
<tr>
<td>(IN)</td>
</tr>
<tr>
<td>V</td>
</tr>
<tr>
<td>O</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sociolinguistic Label</th>
<th>Example error</th>
<th>Target</th>
<th>IPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>(ɪ)-tensing</td>
<td>peaking</td>
<td>picking</td>
<td>/ɪ/ → [i]</td>
</tr>
<tr>
<td>caught/cot merger</td>
<td>com, cot</td>
<td>calm, caught</td>
<td>/ʊ/ → [a], /ʊ/ → [a]</td>
</tr>
<tr>
<td>pre-voiced velar (æ)-raising</td>
<td>beg</td>
<td>bag</td>
<td>/æɡ/ → [eːɡ]</td>
</tr>
<tr>
<td>mistaking (æ) for other Vowel</td>
<td>infect</td>
<td>in fact</td>
<td>/æ/ → [a], /æ/ → [ɛ]</td>
</tr>
<tr>
<td>pre-voiced velar (ɛ)-raising</td>
<td>beg</td>
<td>bake</td>
<td>/ɛɡ/ → [eːɡ]</td>
</tr>
<tr>
<td>(ʌ)-raising</td>
<td>is</td>
<td>us</td>
<td>/ʌ/ → [i], /ʌ/ → [ɪ]</td>
</tr>
<tr>
<td>(ow)-fronting</td>
<td>boot</td>
<td>boat</td>
<td>/ow/ → [u]</td>
</tr>
<tr>
<td>prelateral back vowel merger</td>
<td>full, hole</td>
<td>fool, hull</td>
<td>/ʊl/ ↔ /oʊ/, /uːl/ ↔ /uːl/, /ʌl/ ↔ /oʊ/</td>
</tr>
<tr>
<td>pin/pen merger</td>
<td>pin</td>
<td>pen</td>
<td>/ɪn/ ↔ /ɛn/</td>
</tr>
<tr>
<td>other vowel error</td>
<td>greet</td>
<td>great</td>
<td>varies</td>
</tr>
<tr>
<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)
- (a)
- (æg)
- (æ)
- (æg)
- (o)
- (i)
- (l)
- (d)
- (j)
- (ʔ)
- (TH)
- (ing)

Consonant Errors
- Watch this space!

| (ing) | (TH) | (ʔ) | (i) | (d) | (l) | (æ) | (æg) | (æ) | (eg) | (ow) | (prel) | (IN) | V | O | R | D | NC | NULL | PN | H |
|-------|------|-----|-----|-----|-----|-----|------|-----|------|-----|-------|-----|---|---|---|---|----|-----|----|---|---|---|---|---|---|---|
| 0%    | 0%   | 0%  | 0%  | 0%  | 0%  | 0%  | 0%   | 0%  | 0%   | 0%  | 0%    | 0%  | 9%|36%|16%|4%|1%|4%|14%|16%|
Results

RQ1: Is there a difference in error rates between four ethnicity-related subsamples?

Yes!

• 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>19,142</td>
<td>1.6</td>
</tr>
<tr>
<td>African American</td>
<td>22,773</td>
<td>3.6</td>
</tr>
<tr>
<td>Yakama</td>
<td>22,695</td>
<td>6.3</td>
</tr>
<tr>
<td>ChicanX</td>
<td>9174</td>
<td>6.6</td>
</tr>
</tbody>
</table>

One-Way ANOVA (F(3, 788)=4.514, p<0.001). Tukey’s HSD: Yakama~Caucasian-Am (p=0.04) Caucasian-Am~ChicanX (p=0.00)
#1: Fewest errors \((nf=1.6)\)

What is the by-ethnicity distribution of phonetic error types?

CLOx Errors, by type (Caucasian American Subsample)

### Target
- H: "where’s"
- O: “grading”
- V: “well”
- \((\text{æg})\): “bag”

### Error
- "worse” [wɔ-s]
- “grating”
- “will”
- “beg”

<table>
<thead>
<tr>
<th>(ing)</th>
<th>(TH)</th>
<th>(?)</th>
<th>(u)</th>
<th>(d)</th>
<th>(l)</th>
<th>(ɪ)</th>
<th>(ɔ)</th>
<th>(æ)</th>
<th>(æg)</th>
<th>(ʊ)</th>
<th>(ʃ)</th>
<th>(l)</th>
<th>(d)</th>
<th>(u)</th>
<th>(ʔ)</th>
<th>(TH)</th>
<th>(ing)</th>
</tr>
</thead>
<tbody>
<tr>
<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>9%</td>
<td>36%</td>
</tr>
</tbody>
</table>
CLOx Errors, by type (African American Subsample)

9 additional error types, including:

**Target**
- (V): “head” →
- (IN): “pen” →
- (prel): “Tyler Powell” →
- (ʔ): “dawn” →

**Error**
- “had”
- “pin”
- “taller dollar”
- “done”
CLOx Errors, by type (Yakama Subsample)

4 additional error types

**Target**
(V): “medicine” →
(th): “pen them” →
(ing): “diggin” →

**Error**
“Madison”
“pendam”
“dig in”

### Percentages

<table>
<thead>
<tr>
<th>(ing)</th>
<th>(TH)</th>
<th>(?)</th>
<th>(j)</th>
<th>(d)</th>
<th>(l)</th>
<th>(s)</th>
<th>(æ)</th>
<th>(æg)</th>
<th>(æ)</th>
<th>(ow)</th>
<th>(prel)</th>
<th>(IN)</th>
<th>V</th>
<th>O</th>
<th>R</th>
<th>D</th>
<th>NC</th>
<th>NULL</th>
<th>PN</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>4%</td>
<td>5%</td>
<td>0%</td>
<td>0%</td>
<td>1%</td>
<td>0%</td>
<td>1%</td>
<td>1%</td>
<td>0%</td>
<td>0%</td>
<td>1%</td>
<td>0%</td>
<td>1%</td>
<td>1%</td>
<td>3%</td>
<td>20%</td>
<td>23%</td>
<td>7%</td>
<td>21%</td>
<td>6%</td>
<td>3%</td>
</tr>
</tbody>
</table>
CLOx Errors, by type (Chicanx Subsample)

2 additional consonantal error types

**Target**
- (l): “cheat” ➞ “sheet”
- (?) “a kitten” [əkɪʔn] ➞ “akin”
- (ɛɡ): “peg” ➞ “pig”

### Error Types

<table>
<thead>
<tr>
<th>Error Type</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>(ing)</td>
<td>1%</td>
</tr>
<tr>
<td>(TH)</td>
<td>0%</td>
</tr>
<tr>
<td>(?)</td>
<td>2%</td>
</tr>
<tr>
<td>(d)</td>
<td>3%</td>
</tr>
<tr>
<td>(l)</td>
<td>1%</td>
</tr>
<tr>
<td>(ɪ)</td>
<td>0%</td>
</tr>
<tr>
<td>(æ)</td>
<td>0%</td>
</tr>
<tr>
<td>(æɡ)</td>
<td>0%</td>
</tr>
<tr>
<td>(ʌ)</td>
<td>0%</td>
</tr>
<tr>
<td>(ow)</td>
<td>1%</td>
</tr>
<tr>
<td>(prel)</td>
<td>0%</td>
</tr>
<tr>
<td>(IN)</td>
<td>0%</td>
</tr>
<tr>
<td>V</td>
<td>0%</td>
</tr>
<tr>
<td>O</td>
<td>10%</td>
</tr>
<tr>
<td>D</td>
<td>16%</td>
</tr>
<tr>
<td>R</td>
<td>24%</td>
</tr>
<tr>
<td>NC</td>
<td>5%</td>
</tr>
<tr>
<td>NULL</td>
<td>0%</td>
</tr>
<tr>
<td>PN</td>
<td>9%</td>
</tr>
</tbody>
</table>
By-Task Results

What dialect features appear to be most challenging for our CLOx speech-to-text service (Microsoft)? Are these dialect features more typically found in the more casual speech tasks?

70% of errors were in CS

Mean PER for lexical task materials was relatively low

Figure 2. Errors, by Task. All groups pooled. CS=Conversational Speech, LEX=Lexical Task, RP=Reading Passage
Which sociolinguistic variables were most problematic for the MS ASR system?

- Prelateral merger of /ul/ ~ /ul/ vs. /æ/ vs. /a/
- (th)-stopping /ɔ/ vs. /a/
- CC simplification

Figure 4. PER, by Sociolinguistic variable Class, Task, and Ethnicity.
Conclusions and
Where do we go from here?

This research has accomplished a cross-ethnicity comparison of dialect-based ASR performance

- Important! Quantified contribution of linguistic variables to error profile
- It’s worth it! Eliminate approximately 26% of observed errors
- ASR is a useful tool on the way to “actual” linguistic analysis.

Where does the PNWE team go from here?

- Collaborate on and advocate for leveraging sociolinguistic knowledge of the fine phonetic detail in dialect variation
- Working on new pronunciation model that implements 15 of our targeted sociolinguistic variables
- Building ASR service using freely-available Kaldi architecture
Conclusions and Where do we go from here?

Where can linguists go from here? Some ideas:

• With respect to analysis of sociolectal variation, we need:
  • Further work on *variation* in AAE and other sociolectal varieties
  • Methods for study of multilectal speech
  • More expansive notion of native speaker

• Undoing racial and colonial bias:
  • “Look out for the overlooked”
  • Who gets excluded from linguistic research?
  • Address organizational role-related disparities (employment, tenure and promotion)

“Look out for The Overlooked”
-- folk saying, popularized recently by Kamala Harris in *The Truths We Hold* (2019)
Thank you!

wassink@uw.edu

Perception Test: https://depts.washington.edu/sociolab

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


Koenecke, Allison, Nam, Andrew, Lake, Emily, Nudell, Joe, Quartey, Minnie, Mengesha, Zion, Toups, Connor, Rickford, John R., Jurafsky, Dan, & Goel, Sharad (2020) Racial disparities in automated speech recognition, Proc. of the National Academy of Sciences, 117(14), April 7: 7684-7689.


Reading Passage example

African American (F)

Source: Biadsy et al. (2010)
### Within subsample ANOVA tests of mean difference in PER, by Task

| Group         | Estimate | Std. Error | t value | Pr(>|t|) |
|---------------|----------|------------|---------|----------|
| **African American** |          |            |         |          |
| (Intercept)   | 0.023531 | 0.005538   | 4.249   | 6.59e-05 *** |
| TaskLEX       | -0.012665| 0.007832   | -1.617  | 0.1104   |
| TaskRP        | -0.016262| 0.007832   | -2.076  | 0.0416 * |
| F-statistic: 2.379 on 2 and 69 DF, p-value: 0.1002 |
| **Caucasian American** |          |            |         |          |
| (Intercept)   | 0.027419 | 0.006028   | 4.549   | 2.25e-05 *** |
| TaskLEX       | -0.017608| 0.008525   | -2.065  | 0.04264 * |
| TaskRP        | -0.022984| 0.008525   | -2.696  | 0.00881 ** |
| F-statistic: 3.978 on 2 and 69 DF, p-value: 0.02318 |
| **Yakama**    |          |            |         |          |
| (Intercept)   | 0.032561 | 0.006630   | 4.911   | 5.84e-06 *** |
| TaskLEX       | -0.025233| 0.009376   | -2.691  | 0.00892 ** |
| TaskRP        | -0.030782| 0.009376   | -3.283  | 0.00161 ** |
| F-statistic: 6.124 on 2 and 69 DF, p-value: 0.003562 |
| **Mexican American** |          |            |         |          |
| (Intercept)   | 0.028016 | 0.005270   | 5.316   | 1.23e-06 *** |
| TaskLEX       | -0.017346| 0.007453   | -2.328  | 0.02288 * |
| TaskRP        | -0.025036| 0.007453   | -3.359  | 0.00128 ** |
| F-statistic: 5.922 on 2 and 69 DF, p-value: 0.004229 |