Racial Disparities in Automated Speech Recognition

Allison Koenecke
Cornell Information Science
Nov 15, 2023

PNAS Proceedings of the National Academy of Sciences of the United States of America

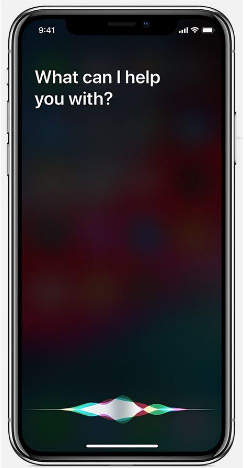
RESEARCH ARTICLE

Racial disparities in automated speech recognition

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel

PNAS April 7, 2020 117 (14) 7684-7689; first published March 23, 2020; https://doi.org/10.1073/pnas.1915768117

Automated Speech Recognition (ASR)



Why do ASRs matter?

- Applications in:
  - Digital device interaction for individuals with physical impairments

(Images from real services sold by: tecla, BMW, Olympus America, and Planet Depos)

Why do ASRs matter?

- Applications in:
  - Digital device interaction for individuals with physical impairments
  - Car systems for safer driving

(Images from real services sold by: tecla, BMW, Olympus America, and Planet Depos)

Why do ASRs matter?

- Applications in:
  - Digital device interaction for individuals with physical impairments
  - Car systems for safer driving
  - Medical dictation devices for doctors recording patient notes

(Images from real services sold by: tecla, BMW, Olympus America, and Planet Depos)




Why do ASRs matter?

- Applications in:
  - Digital device interaction for individuals with physical impairments
  - Car systems for safer driving
  - Medical dictation devices for doctors recording patient notes
  - Court transcription services

(Images from real services sold by: tecla, BMW, Olympus America, and Planet Depos)

Why do ASRs matter?

- Applications in:
  - Digital device interaction for individuals with physical impairments
  - Car systems for safer driving
  - Medical dictation devices for doctors recording patient notes
  - Court transcription services
- Downstream impacts

(Images from real services sold by: tecla, BMW, Olympus America, and Planet Depos)

Who are we auditing?

Audits in analogous domains

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *FAT*.

Audio Data

- We use two compilations of sociolinguistic interviews:
  - Corpus of Regional African American Language (CORAAL)
  - Voices of California (VOC)

Audio Data

- We use two compilations of sociolinguistic interviews:
  - Corpus of Regional African American Language (CORAAL)
  - Voices of California (VOC)

https://www.unh.edu/inquiryjournal/spring-2015/do-you-%E2%80%9Cpark-your-car%E2%80%9D-or-%E2%80%9Cpahk-your-cah%E2%80%9D-changing-dialect-southern-new-hampshire

Audio Data

- We use two compilations of sociolinguistic interviews:
  - Corpus of Regional African American Language (CORAAL)
  - Voices of California (VOC)

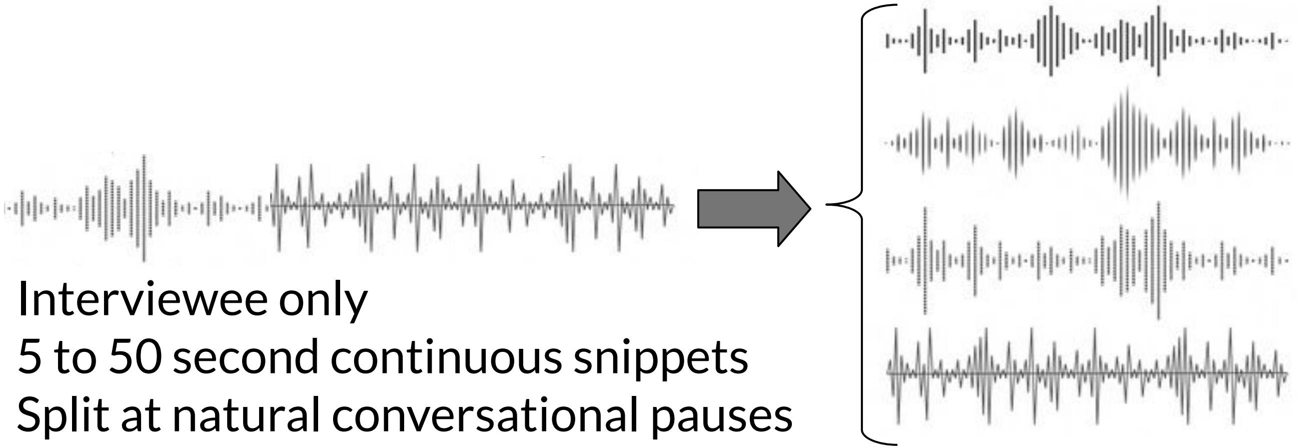
Audio Data

- We use two compilations of sociolinguistic interviews:
  - Corpus of Regional African American Language (CORAAL)
  - Voices of California (VOC)
- Advantage: unseen data to audit black-box ASR systems
  - Else, it may already be used as training data

Audio Data

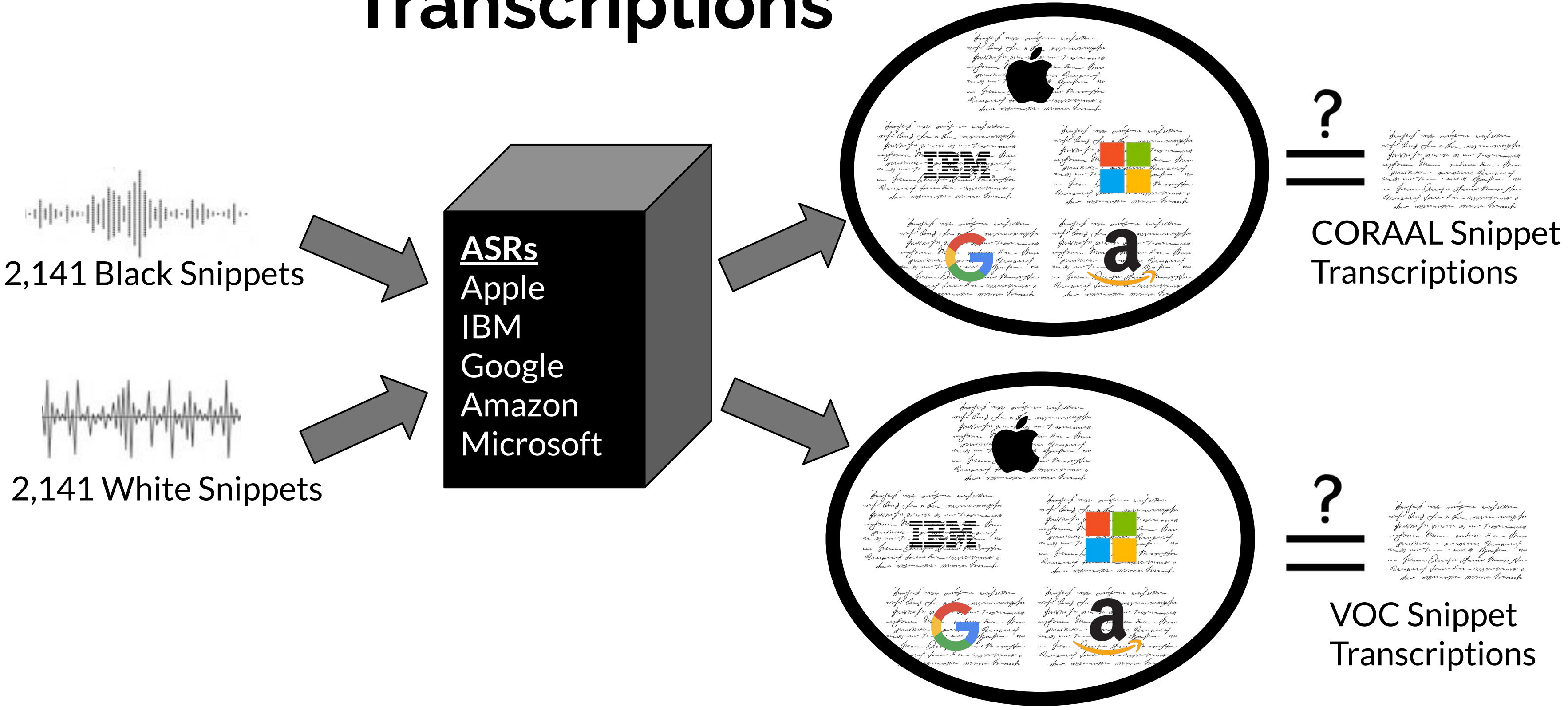
- We use two compilations of sociolinguistic interviews:
  - Corpus of Regional African American Language (CORAAL)
  - Voices of California (VOC)
- Advantage: unseen data to audit black-box ASR systems
  - Else, it may already be used as training data
- Both sources yield ~40 hours of interviewee speech and human-generated ground-truth transcripts

Audio Processing



- Interviewee only
- 5 to 50 second continuous snippets
- Split at natural conversational pauses
- Propensity match on age, gender, duration

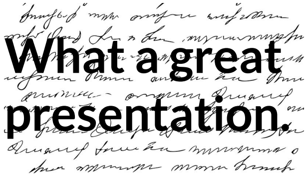
Transcriptions



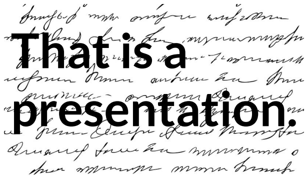
Word Error Rate

$$\text{WER} = \frac{\text{Substitutions} + \text{Deletions} + \text{Insertions}}{\#\ \text{Ground Truth Words}}$$

Ground Truth:



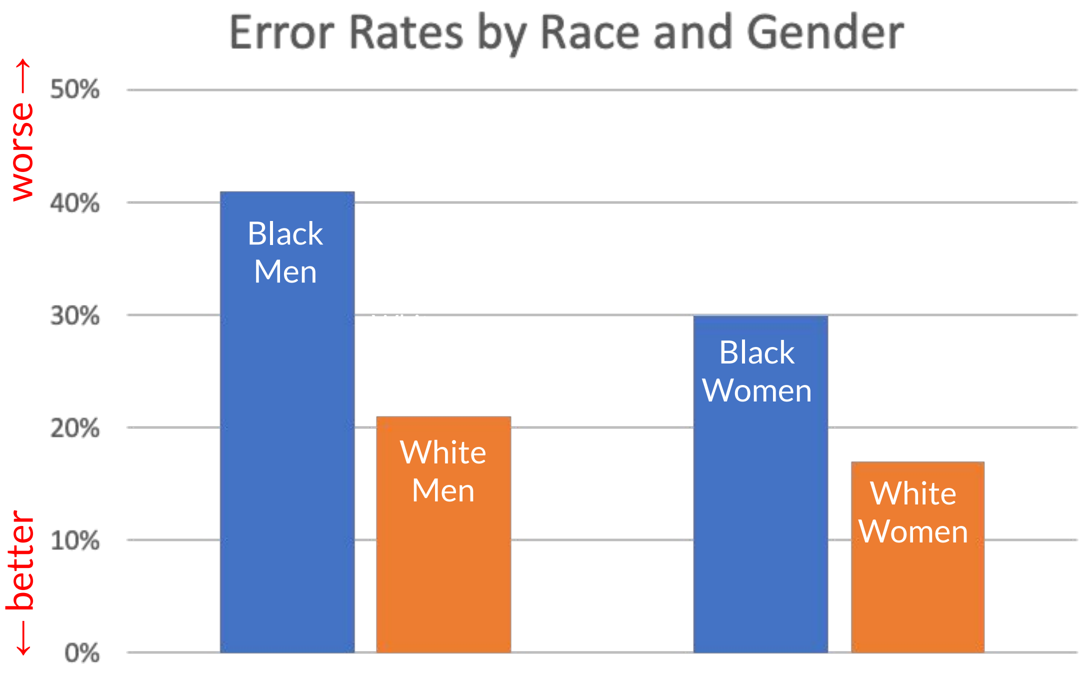
Transcription:



~~That~~ ~~is~~ a great presentation.
What

WER = 3 / 4 = **0.75**

Black WER are ~2x White WER



White Man Sample

WER = 0.21

Well , when I was ~~that's~~ when I was really young I had a book of basketball statistics ~~.~~ ~~No~~ and I would spend a lot of time a lot of time reading them . And for some reason , I forget why now , but Jason Kidd ~~pain.~~ ended ~~Be~~ up being my favorite player .

Black Woman Sample

WER = 0.30

And these little snow
~~If~~ she had ~~a~~ ~~photo~~ like ~~no~~ cone

I but
things. ~~Don~~ don't even know what it was, ~~does~~

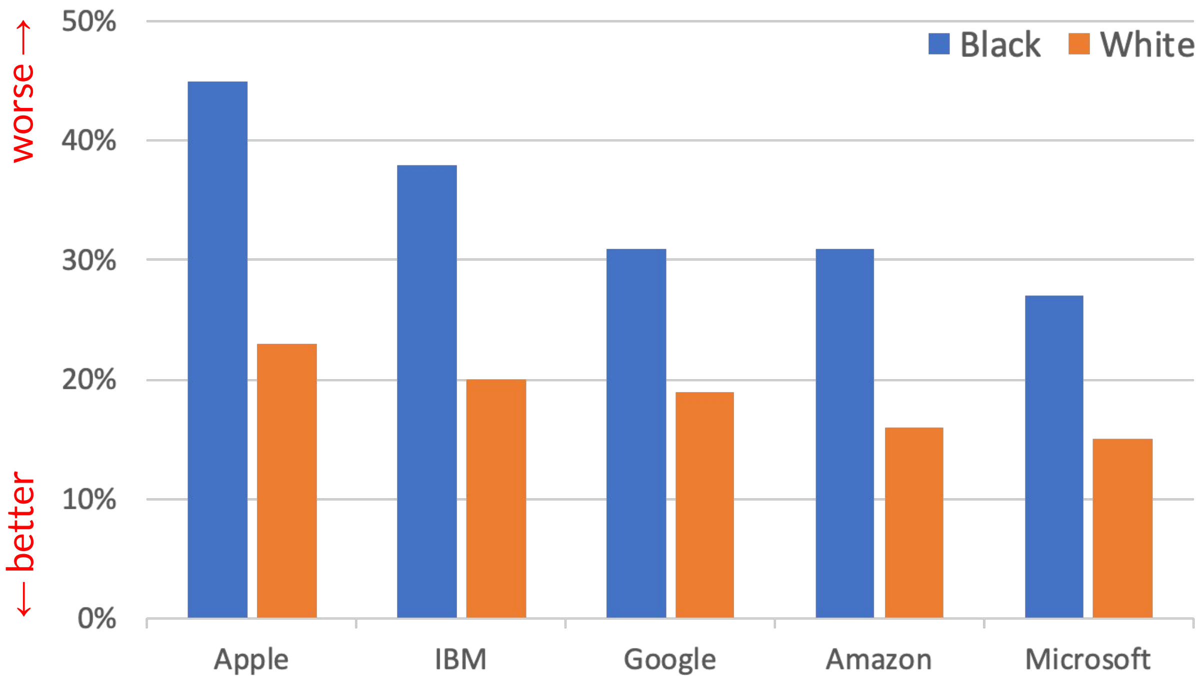
snow cones
it's not like the ~~smell~~ ~~comes~~ up here. Like,

I don't know how to explain it, but you know

- ~~a~~ bag of candy for a quarter. Like, a full

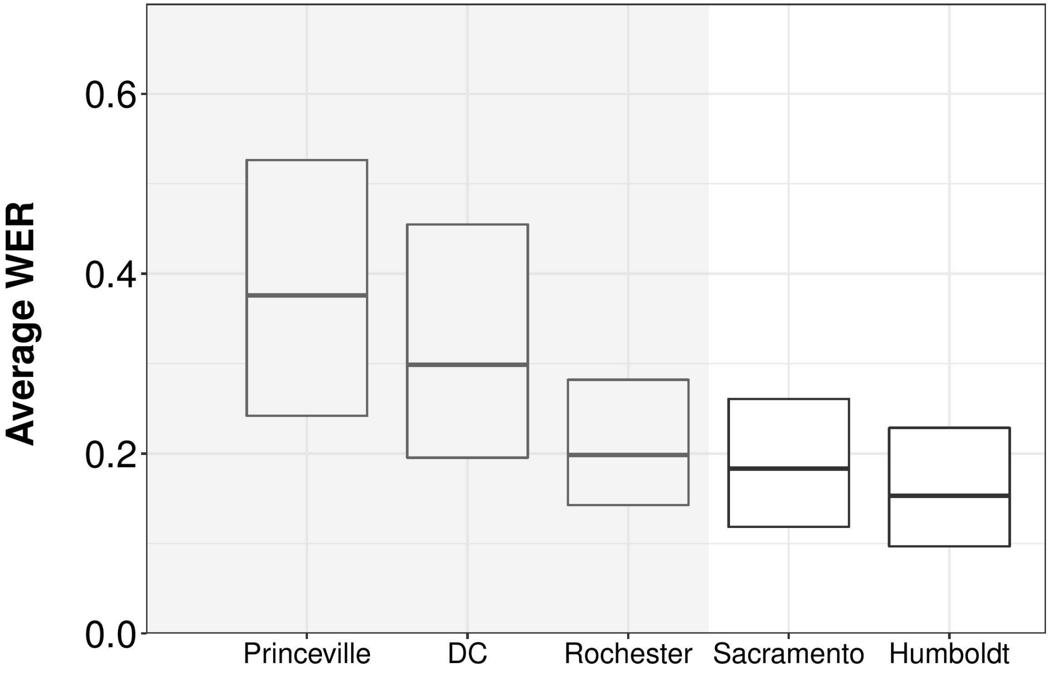
bag of candy for a quarter.

Errors consistent across firms



How do we know these are racial disparities?

High geographic variation in WER



On "AAVE" and "SE"

- Linguists use "vernacular" to distinguish varieties with particular researched features, as against the varieties that "all" African Americans use (e.g. AAL / AAE)
    - "Language and linguistics on trial: Hearing Rachel Jeantel (and other vernacular speakers) in the courtroom and beyond" (Rickford & King, 2016)
    - "Spoken Soul: The Story of Black English" (Rickford & Rickford, 2000)
    - "Suite for Ebony and Phonics" (Rickford, 1997)

- We use the term "Standard," but only referring to regularization of features and not desirability

Dialect Density Measure

- African American Vernacular English is spoken by nearly 12% of all Americans

Dialect Density Measure

- African American Vernacular English is spoken by nearly 12% of all Americans
- Count hand-coded AAVE linguistic features in random sample of 50 snippets per interview site

Dialect Density Measure

- African American Vernacular English is spoken by nearly 12% of all Americans
- Count hand-coded AAVE linguistic features in random sample of 50 snippets per interview site
- Grammatical and phonological examples:
    - **Zero copula:** They gone
    - **Future *be*:** He be here tomorrow
    - **Final consonant cluster reduction:** band → ban’
    - **Hapology:** mississippi → misipi

Fewer AAVE Features

WER = 0.03

Grow
~~World~~ older , we get darker . So I was extremely light when I was a child and very skinny . And so I was like an outcast because I was made fun of because I was the white girl at the school .

More AAVE Features

WER = 0.56

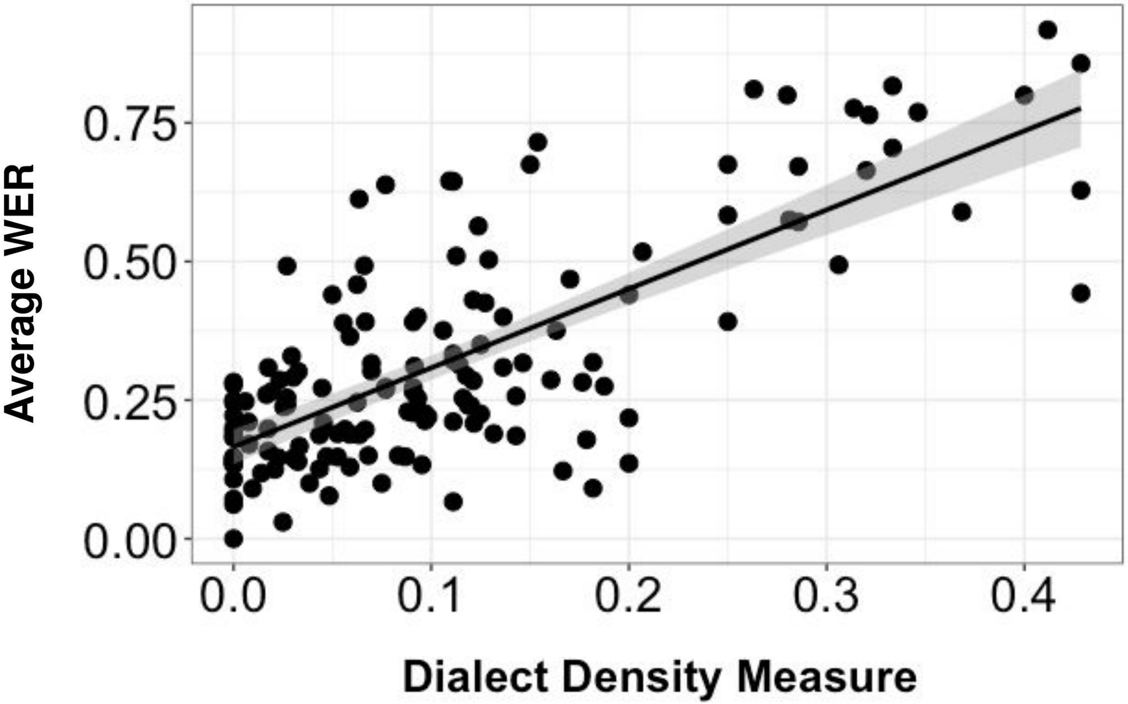
In second grade teacher gave
~~With~~ ~~seven~~ ~~braids.~~, ~~He'd~~ ~~give~~ me ~~a~~ nickname

sneaky
Snake cause well she said I was ~~sneaking~~. You

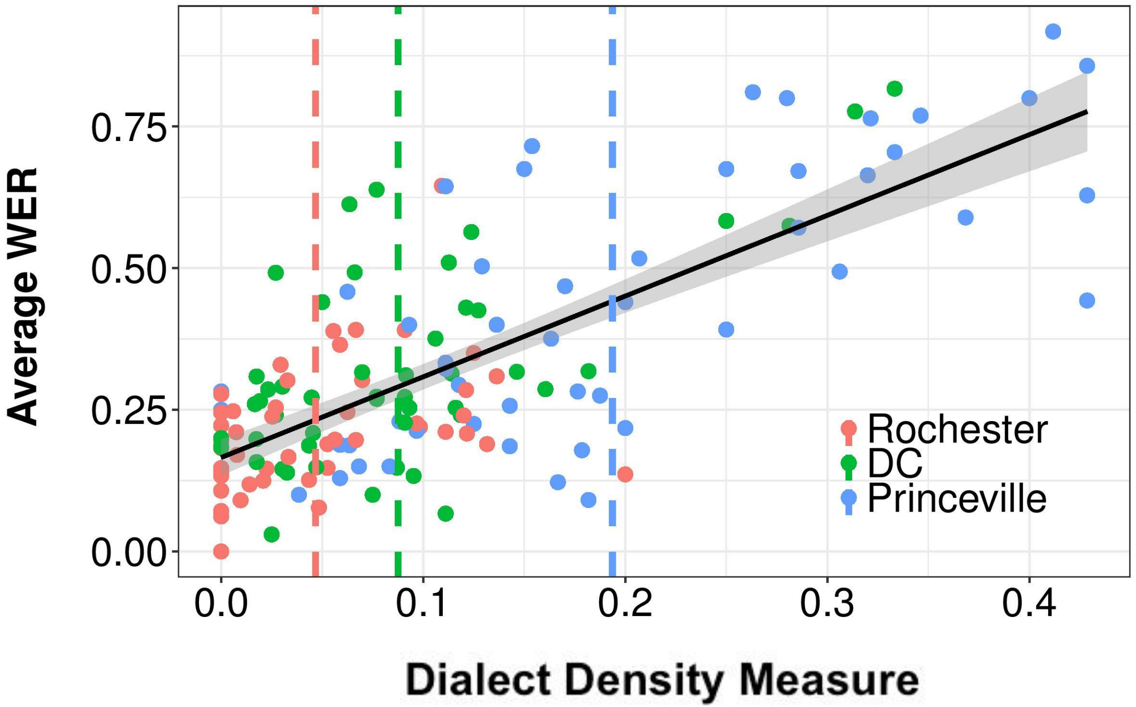
I be
know. ~~Me~~ ~~sit~~ sitting in one place and she

turn around sitting
~~China,~~ ~~man,~~ I'm ~~staying~~ someplace else.

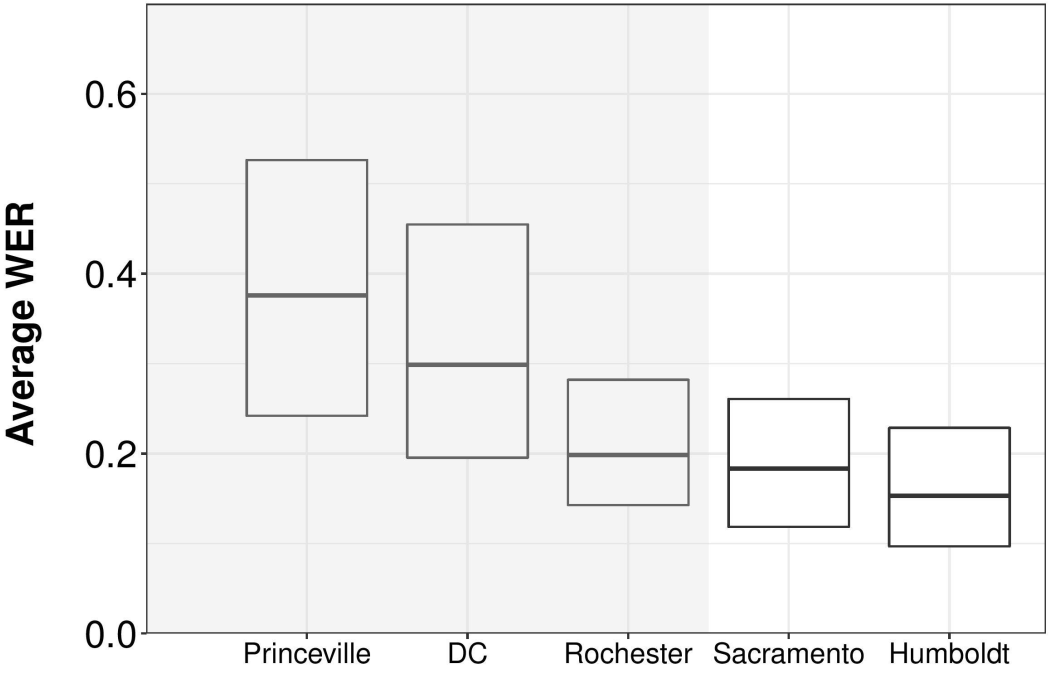
Positive correlation of DDM and WER



Positive correlation of DDM and WER



High geographic variation in WER



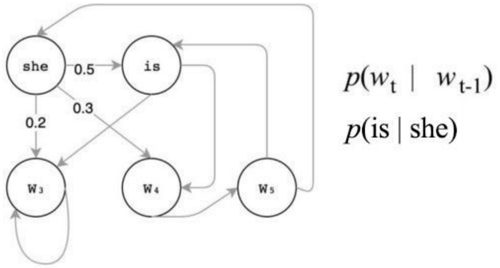
“Gender and Dialect Bias in YouTube’s Automatic Captions” (Tatman, 2017)

Why do ASRs yield these racial disparities?

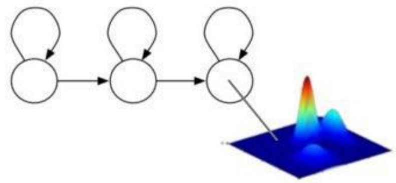
Why do ASRs perform poorly on AAVE?

Modern ASRs have two underlying components that could result in the racial disparity we see in performance:

1. Language models



2. Acoustic models



Why do ASRs perform poorly on AAVE?

Modern ASRs have two underlying components that could result in the racial disparity we see in performance:

1. Language models
   - **Test 1: Lexicon**
   - **Test 2: Grammar**
2. Acoustic models
   - **Test 3: Phonology**

Why do ASRs perform poorly on AAVE?

Modern ASRs have two underlying components that could result in the racial disparity we see in performance:

1. Language models
   - ❌ Test 1: Lexicon
   - ❌ Test 2: Grammar
2. Acoustic models
   - Test 3: Phonology

Acoustic Model Test

- Find Black and white speakers saying identical phrases in our sample

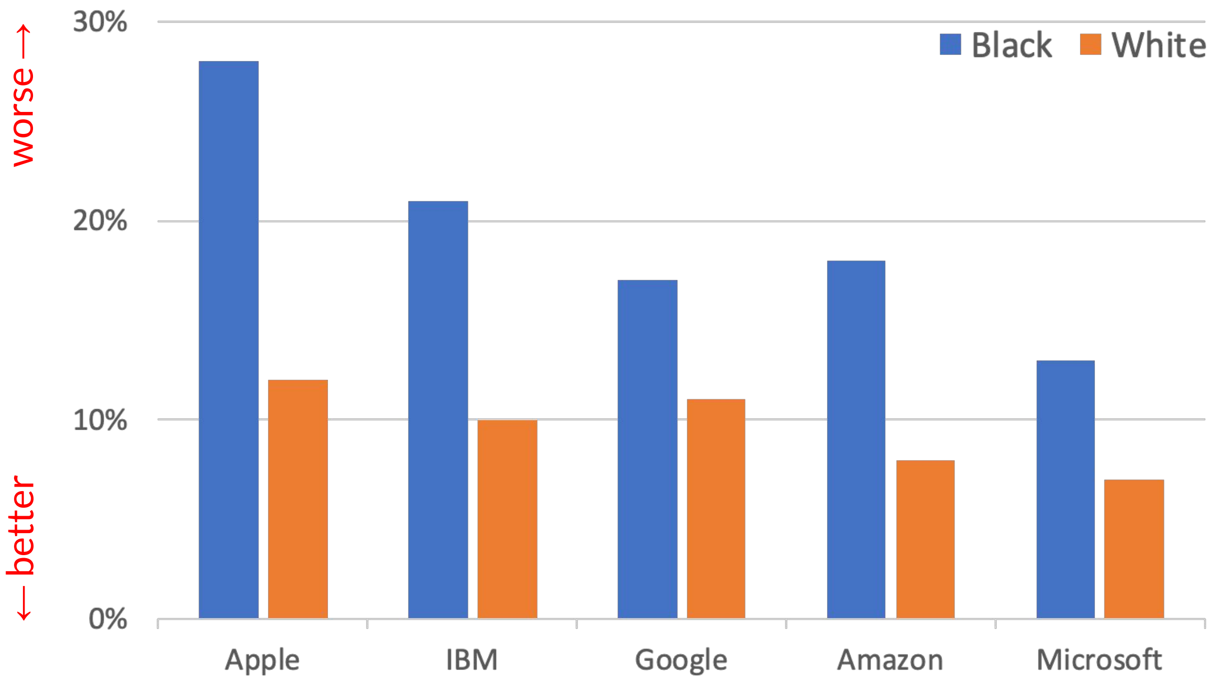
Acoustic Model Test

- Find Black and white speakers saying identical phrases in our sample
- Match pairs of Black and white speakers (of the same gender and similar age) uttering 5 to 8 word n-grams
    - *“and then a lot of the”*
    - *“and my mother was a”*

Acoustic Model Test

- Find Black and white speakers saying identical phrases in our sample
- Match pairs of Black and white speakers (of the same gender and similar age) uttering 5 to 8 word n-grams
  - *“and then a lot of the”*
  - *“and my mother was a”*
- Compare error rates across the 206 matched phrases

Black WER ~2x White WER, again



Why do ASRs perform poorly on AAVE?

Modern ASRs have two underlying components that could result in the racial disparity we see in performance:

1. Language models — ✗ Test 1: Lexicon — ✗ Test 2: Grammar
2. Acoustic models — ✓ Test 3: Phonology

Our study showed...

1. All five ASR systems exhibited substantial racial disparities as measured by average WER
   a. 0.35 for Black speakers, 0.19 for white speakers
2. Racial disparities in ASR performance are traced to the acoustic model
   a. Related to racial differences in rhythm, pitch, syllable accenting, vowel duration, lenition

Call to action

- More diverse data should be collected: both of AAVE speech, and other non-standard varieties of English

Call to action

- More diverse data should be collected: both of AAVE speech, and other non-standard varieties of English

RESEARCH-ARTICLE OPEN ACCESS

Augmented Datasheets for Speech Datasets and Ethical Decision-Making

Authors: Orestis Papakyriakopoulos, Anna Seo Gyeong Choi, William Thong, Dora Zhao, Jerone Andrews, Rebecca Bourke, Alice Xiang, Allison Koenecke Authors Info & Claims

FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency • June 2023 • Pages 881–904
• https://doi.org/10.1145/3593013.3594049

Call to action

- More diverse data should be collected: both of AAVE speech, and other non-standard varieties of English
- The speech recognition community needs to invest resources to ensure ASR systems -- and the institutions that build them -- are broadly inclusive

Call to action

- More diverse data should be collected: both of AAVE speech, and other non-standard varieties of English
- The speech recognition community needs to invest resources to ensure ASR systems -- and the institutions that build them -- are broadly inclusive
- ASR developers should regularly assess and publicly report progress over time

Call to action

- More diverse data should be collected: both of AAVE speech, and other non-standard varieties of English
- The speech recognition community needs to invest resources to ensure ASR systems -- and the institutions that build them -- are broadly inclusive
- ASR developers should regularly assess and publicly report progress over time
- Learn from algorithmic & legislative progress made in other domains (e.g., computer vision)

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *FAT*.

Progress?

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *FAT*.

United States House Committee on Oversight and Government Reform

May 22, 2019

Hearing on

Facial Recognition Technology (Part 1): Its Impact on our Civil Rights and Liberties

Big tech companies back away from selling facial recognition to police. That's progress.

After IBM, Amazon, and Microsoft upend their facial recognition businesses, attention turns to federal lawmakers.

By Rebecca Heilweil | Updated Jun 11, 2020, 5:02pm EDT

ASR Progress?

Federal Communications Commission

FCC Seeks Comment on Petition Regarding Live Captioning Quality Metrics and Use of Automated Speech Recognition

On August 14, 2019, the FCC's Consumer and Governmental Affairs Bureau released a Public Notice inviting public comment on a petition for declaratory ruling and rulemaking filed by a coalition of consumer and academic organizations in regard to live captioning quality metrics and the use of automated speech recognition techniques.

ASR Progress?

Federal Communications Commission

FCC Seeks Comment on Petition Regarding Live Captioning Quality Metrics and Use of Automated Speech Recognition

On August 14, 2019, the FCC's Consumer and Governmental Affairs Bureau released a Pu
declaratory ruling and rulemaking filed by a coalition of consumer and academic organiz
the use of automated speech recognition techniques.

Jul 6, 2023 - Technology

NYC law promises to regulate AI in hiring, but leaves crucial gaps

Ivana Saric

Questions?

→ koenecke@cornell.edu

→ fairspeech.stanford.edu