Fair or Fare? Understanding Automated Transcription Error Bias in Social Media and Videoconferencing Platforms
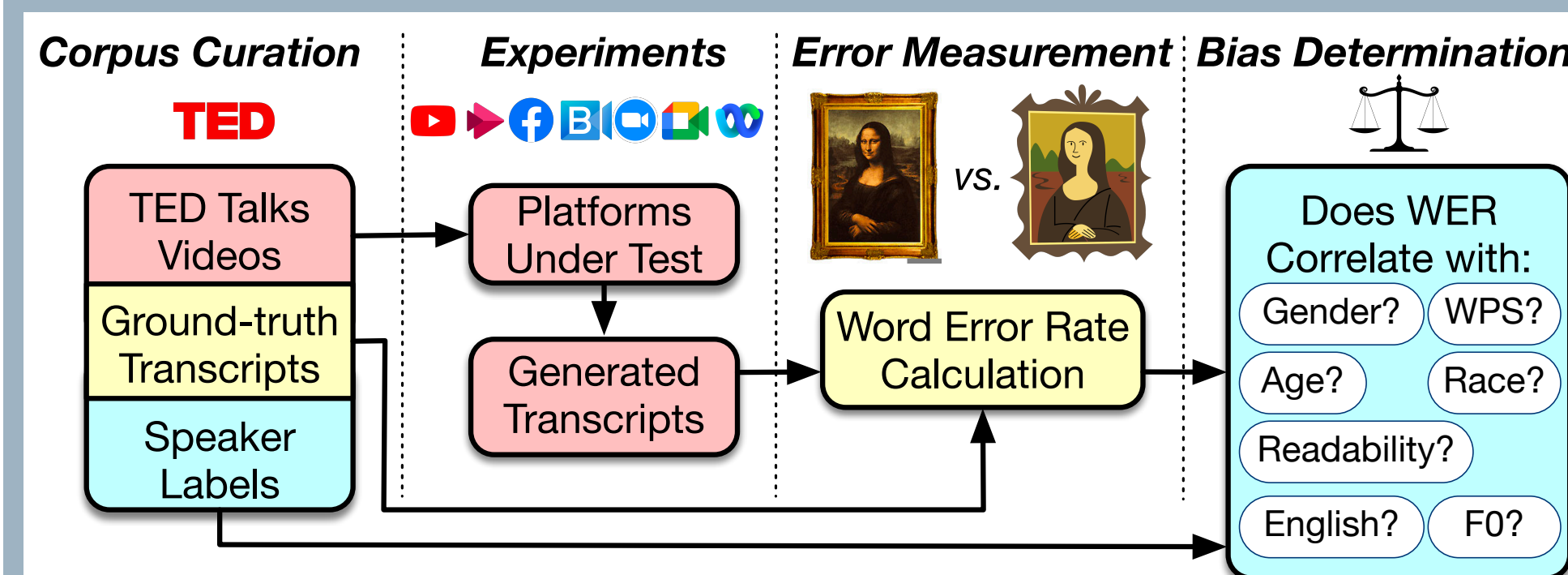
Daniel J. Dubois[1], Nicole Holliday[2], Kaveh Waddell[3,4], David Choffnes[1]
[1]Northeastern University [2]Pomona College [3]Consumer Reports [4]Stanford University

Northeastern University Khoury College of Computer Sciences · Pomona College · Consumer Reports · Stanford University

MOTIVATION

- **Context:** Automated transcription is important in Social Media and Videoconferencing Platforms.
- **Examples**: Accessibility, remote learning, remote working, social activities, etc.
- **Problem**: Platforms may exhibit systematic bias in terms of transcription errors.

METHODOLOGY

Corpus Curation — TED: TED Talks Videos, Ground-truth Transcripts, Speaker Labels → *Experiments*: Platforms Under Test, Generated Transcripts → *Error Measurement*: Word Error Rate Calculation → *Bias Determination*: Does WER Correlate with: Gender? WPS? Age? Race? Readability? English? F0?

- **MonoTED corpus curation:** web scraping and perception
- **Speaker characteristics**: gender, age, language, race, voice frequency
- **Speech characteristics**: speech rate, speech readability
- **Experiments:** 3 social media platforms, 4 videoconferencing platforms
- **Transcription error metric**: Word Error Rate (WER)
- **Bias determination**:
  WER analysis by speaker/speech characteristics
  Linear Mixed Effects Regression (LMER) model
  Analysis of Variance (ANOVA) testing

$$WER = \frac{S+D+I}{N}$$

S: number of word substitutions
D: number of word deletions
I: number of word insertions
N: number of words in the ground truth transcript

RESEARCH QUESTIONS

- **Are certain groups of speakers subject to more transcription errors?**
- **What factors correlate with these biases?**

PLATFORMS

- **Social Media:** YouTube, Microsoft Stream, Facebook
- **Videoconferencing**: BlueJeans, Zoom, Meet, Webex

NEW "MonoTED" CORPUS

- TED talks
- Monologues
- 1.8M words
- 194 hours
- 846 speakers
- Ground truth transcripts
- New labels ➤

Characteristic	Value	# of speakers
Perceived gender	Male	548 (65%)
	Female	298 (35%)
English language	First (US)	478 (56%)
	First (non-US)	216 (26%)
	Second	152 (18%)
Perceived race	White	651 (77%)
	Black	77 (9%)
	Asian	104 (12%)
	Latino	14 (2%)

RESULTS

We found **higher transcription error rates** for the male gender, non-native English speakers, speech rates (words per second) lower and higher than the median, and speech that is easier to read than the median.
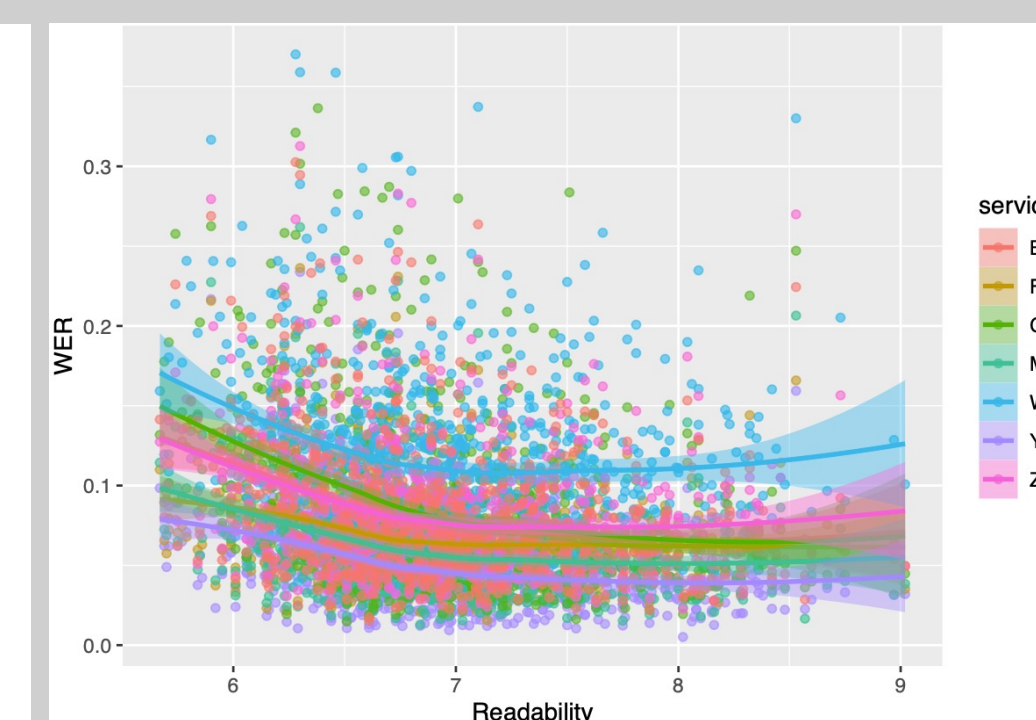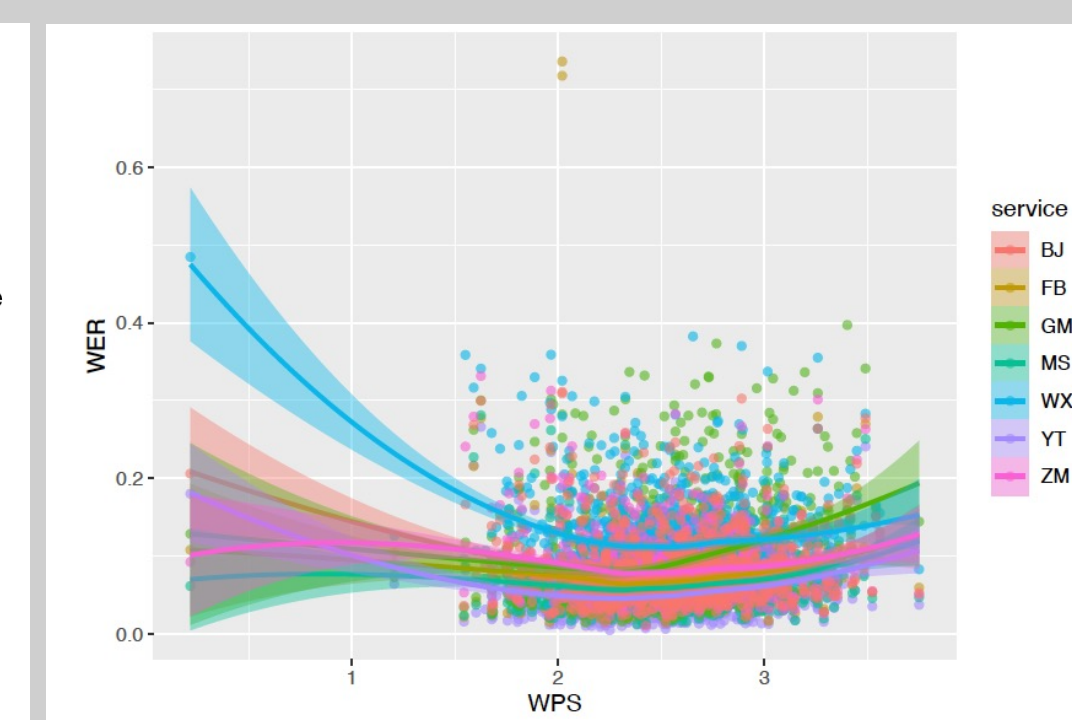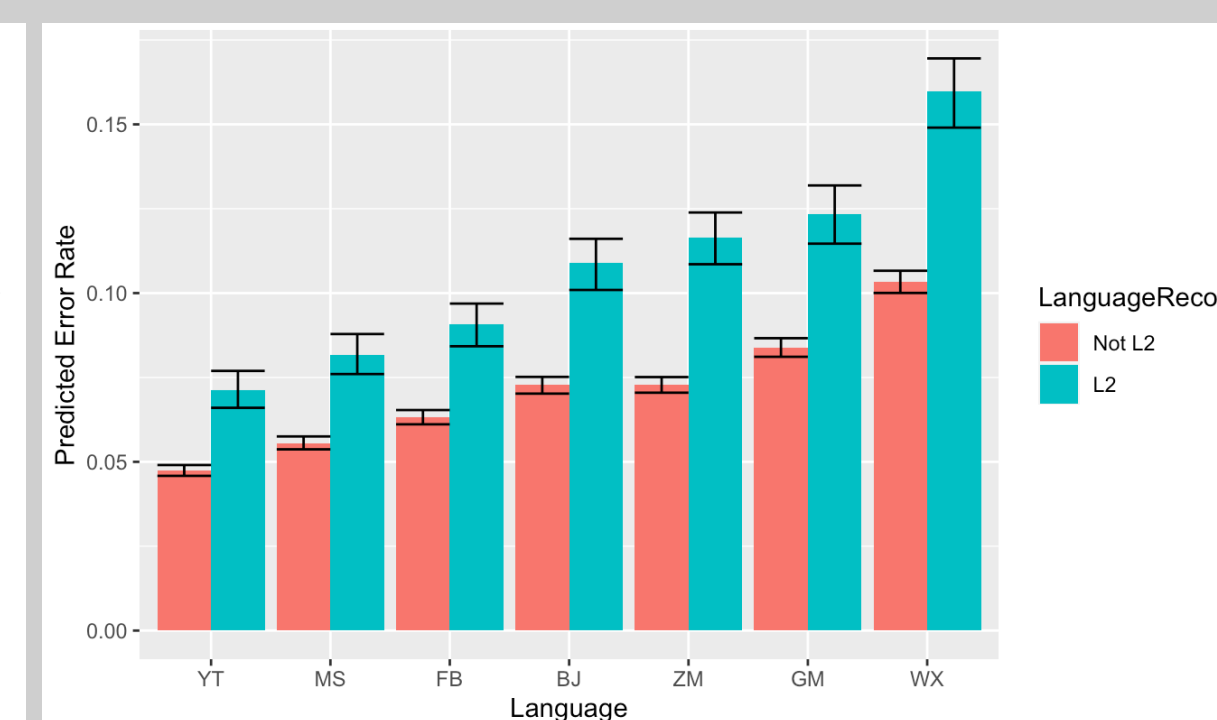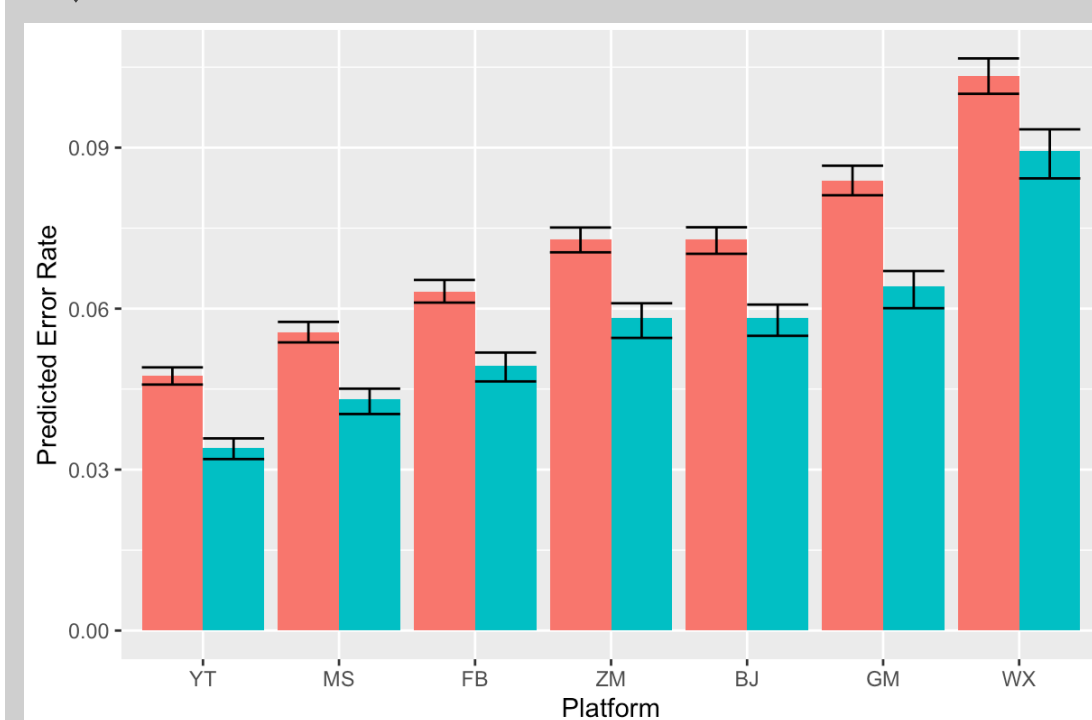
No statistically significant bias for speakers based on race, fundamental frequency of their voice (F_0), and age.

CORPUS AND DATASET

- https://github.com/NEU-SNS/MonoTED
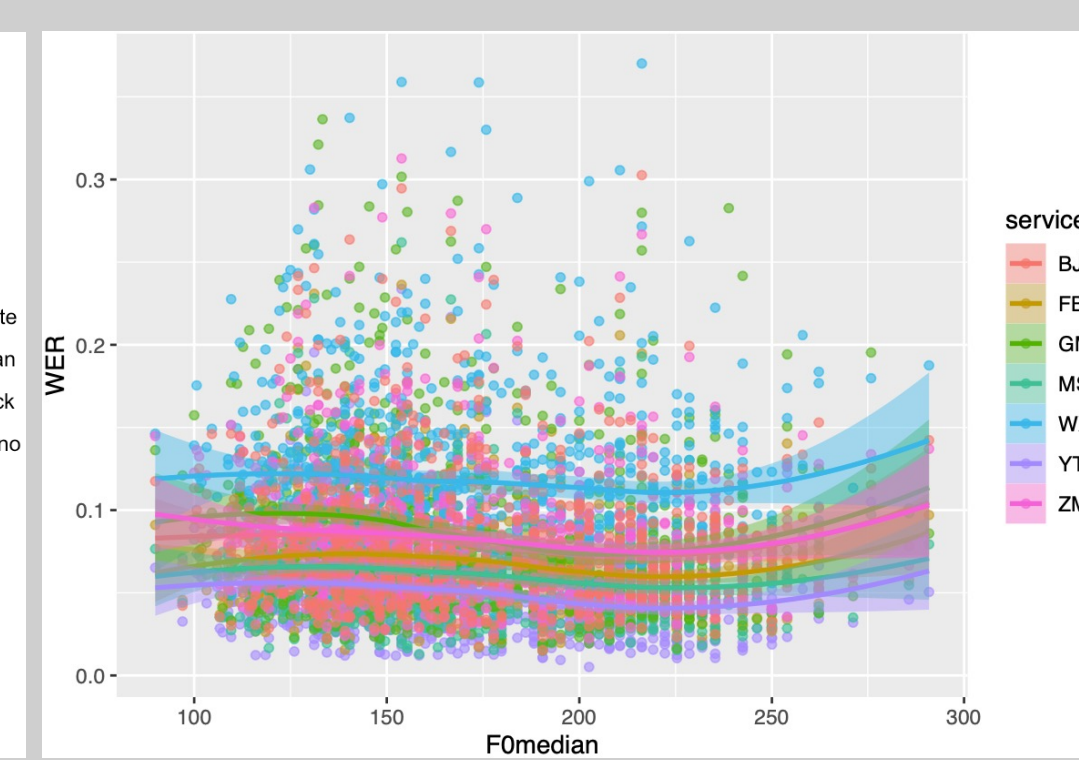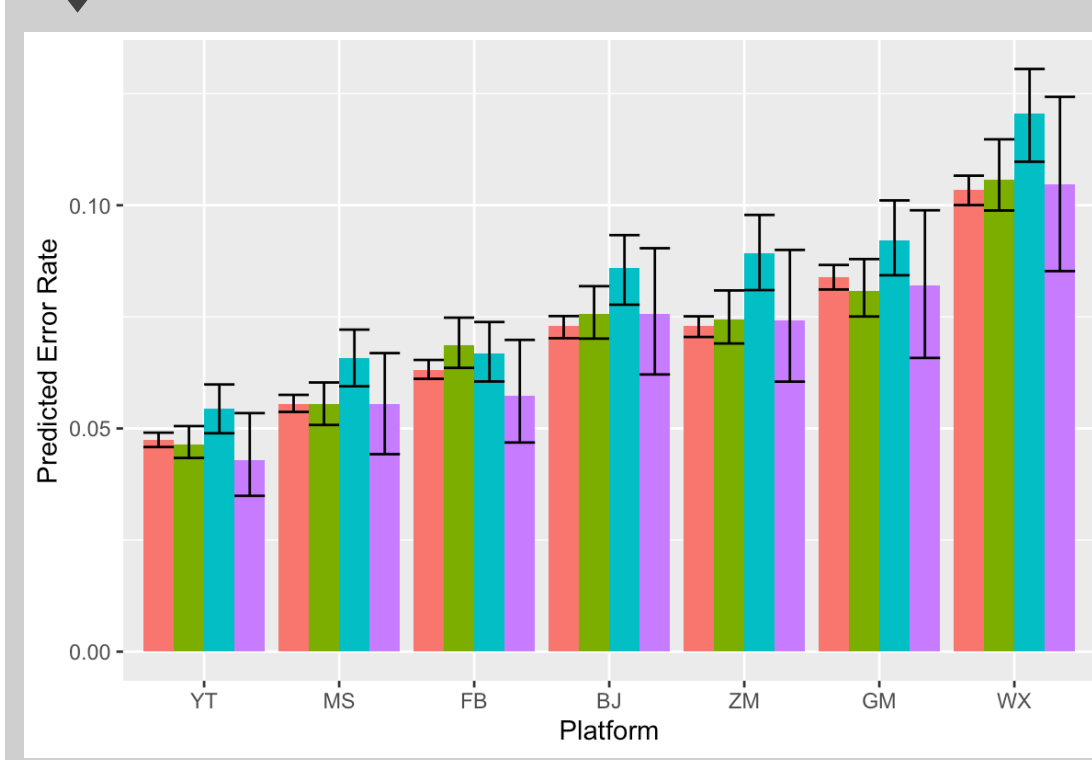
TAKEAWAYS

- **Our results confirm previous findings on biases in generic voice recognition systems, except for race.**
  Possible explanation for race: TED talks are highly-rehearsed speeches.
- **Videoconferencing platforms show higher errors than social media ones.**
  Possible explanation: videoconferencing transcripts are produced in real time.
- **Voice frequency (F_0) does not explain gender bias.**
- **Possible bias mitigation**: platform retraining with more representative samples.