Overlearning speaker race in sociolinguistic auto-coding

Dan Villarreal, Department of Linguistics, Pitt

Motivation

- **Coding**, categorizing linguistic options (e.g., *car* vs "*cah*"), is an important but time-intensive step in socioling research
- Villarreal et al. (2020) used machine learning (random forests) to automate coding based on sound properties
  - Auto-codes matched listener judgments (Fig. 1)
- Other AI applications perform worse for Black than White individuals—what about this auto-coding algorithm?

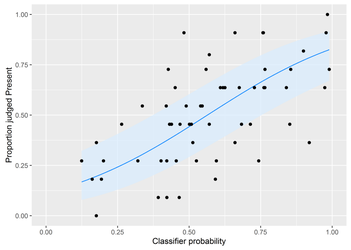

Figure 1: Auto-coder's estimated probability that (r) tokens were Present (e.g., *car*) compared to the proportion of 11 trained listeners who judged (r) tokens as Present; line and 95% confidence band from mixed-effects model of judgments (Villarreal et al. 2020)

Project Description

- Data: ~11,000 tokens of (r) (e.g., *car* vs "*cah*") from Black and White speakers of New England English
- Procedure: Run auto-coders with different unfairness mitigation strategies
- Goal: Assess how these strategies affect fairness (disparity in coding accuracy)

Context

- In domains like criminal justice (Angwin et al. 2016) and ASR (Koenecke et al. 2020), algorithms tend to perform worse on Black than White individuals
  - AI fairness is inherently in tension with performance (Kleinberg et al. 2017)
- These investigations tend to happen after algorithms are in wide use, making AI fairness an afterthought

University of Pittsburgh

Does an algorithm that codes linguistic data **perform differently for Black and White speakers?**

Project Deliverables

- Expand our understanding of the limitations of sociolinguistic auto-coding
- Open up new avenues of research into how intergroup acoustic differences translate to auto-coding performance
- Data preparation complete by August 2021, analysis by January 2022, submission to *Linguistics Vanguard* by April 2022
- Next step: Apply for NSF Fairness in AI grant in summer 2022

Potential Impact

- Introduces AI fairness to a new algorithm in its infancy rather than waiting until it is in wide use
  - Interrupt trend by which new AI methods increase and reproduce racial injustice
- Broaden AI fairness research to a domain with different stakes
- Increase viability of a time-saving method for sociolinguistic research

References

Angwin, Julia, Jeff Larson, Surya Mattu & Lauren Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. In *ProPublica*.

Kleinberg, Jon, Sendhil Mullainathan & Manish Raghavan. 2017. Inherent trade-offs in the fair determination of risk scores. Paper presented to the 8th Innovations in Theoretical Computer Science Conference, Dagstuhl, Germany, 2017.

Koenecke, Allison, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky & Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*.201915768.

Villarreal, Dan, Lynn Clark, Jennifer Hay & Kevin Watson. 2020. From categories to gradience: Auto-coding sociophonetic variation with random forests. *Laboratory Phonology* 11.1-31.

Acknowledgments

Thanks to Monica Nesbitt, Jim Stanford, Lynn Clark, Jen Hay, Kevin Watson, Vica Papp, and all our speakers for making this project possible