Overlearning speaker gender in sociolinguistic auto-coding
Metrics and remedies

Dan Villarreal

Department of Linguistics, University of Pittsburgh; d.vill@pitt.edu

University of Pittsburgh

Background

- In **sociolinguistic auto-coding (SLAC)**, machine learning is used to assign variants to tokens of variables based on acoustic features [1–3]
- Research on **AI fairness** has found that predictive algorithms can reproduce intergroup biases in the data they're trained on [e.g., 4]
  - There are multiple ways to define/measure AI fairness, and it's mathematically proven that they're mutually exclusive [e.g., 5]
  - Fortunately, several strategies exist to mitigate unfairness
- It's possible that SLAC may make predictions about variants based not on legitimate cues to variant identity, but inadvertently on group membership
  - This would be highly problematic, given the central importance in sociolinguistics of correlating speaker groups to differences in variable usage

Research questions

1. Which fairness metric(s) are appropriate for SLAC?
2. Is SLAC prone to differential predictions by speaker group?
3. How can unfairness be mitigated in SLAC?

In this project, I look at **auto-coding English non-prevocalic /r/** (Absent vs. Present) and fairness with respect to **speaker gender**.

RQ1: Defining fairness for SLAC

- Confusion matrix from /r/ auto-coder in [3, 6] has True Absent, False Absent, False Present, & True Present
- *Overall accuracy* = (TA + TP) / (TA + FA + FP + TP)
- Absent *class accuracy* = TA / (TA + FA)
- Present *class accuracy* = TP / (TP + FP)

		Actual	
		Absent	Present
Predicted	Abs	3137	522
	Pres	247	783

Among the fairness criteria defined by [5]…

Makes sense for SLAC: overall accuracy equality (*OAE*)

We want the /r/ auto-coder to code women & men equally well, regardless of whether tokens are Absent or Present

Makes sense for SLAC: class accuracy equality (*CAE*)

We want the /r/ auto-coder to code women's and men's Absent tokens equally well, and their Present tokens equally well

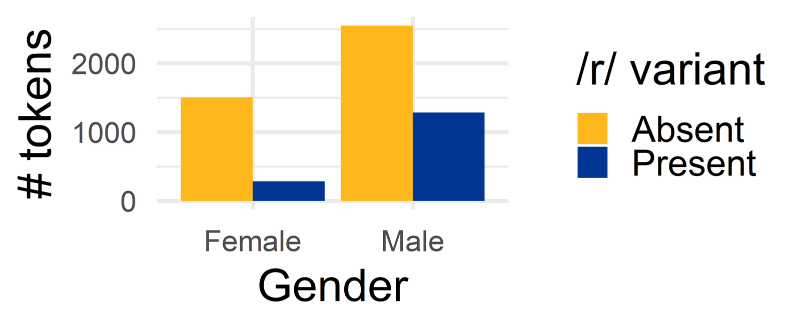
Doesn't make sense for SLAC: statistical parity

We do **not** want the /r/ auto-coder to predict that women & men are equally rhotic

RQ2: Fairness assessment

Gender fairness assessed for Southland New Zealand English /r/ auto-coder in [3, 6]

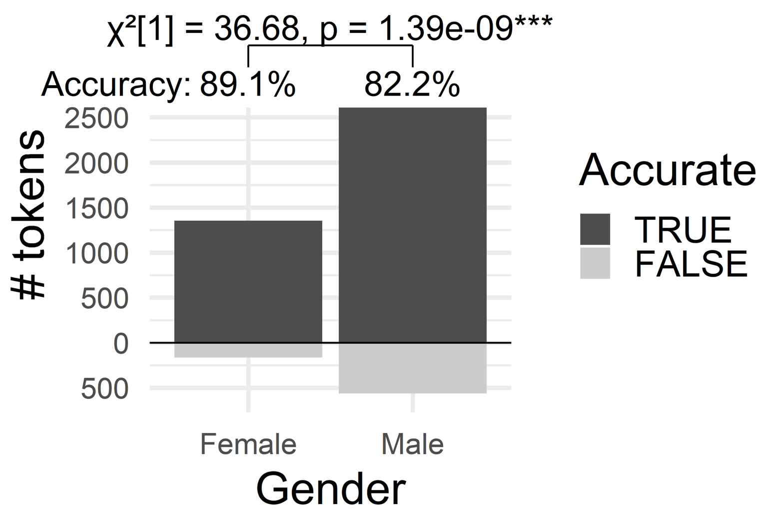
- 5620 hand-coded tokens
  - Male /r/s outnumber female 2:1
  - Male /r/s signif more rhotic
- Trained on 180 acoustic measures (formants, pitch, intensity, timing)
- Auto-coders implemented as random forest in R using `caret` and `ranger` [7–9]
  - Optimized for performance, not fairness



Overall accuracy equality: unfair

Women's /r/s auto-coded with significantly **greater overall accuracy** than men's

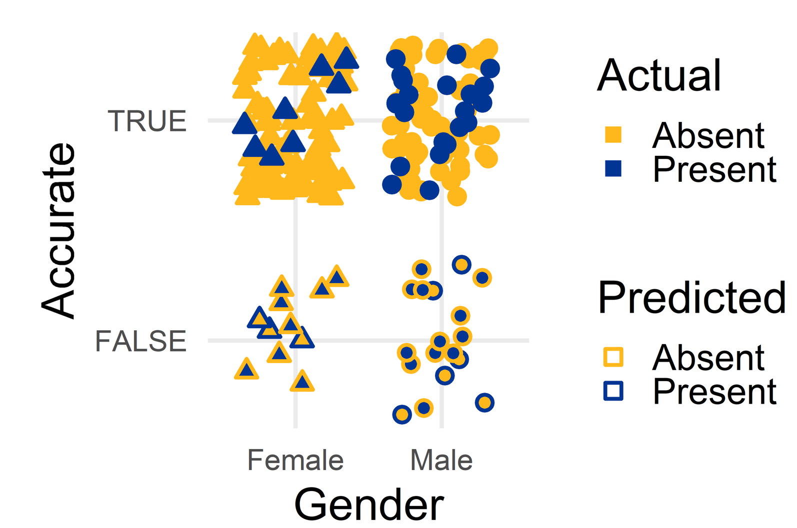
- Men have worse overall accuracy despite a training set twice as large
  - Size of training set doesn't guarantee good auto-coding performance



Class accuracy equality: unfair

Class accuracies **unequal** across gender

- Absent /r/s coded better when speaker is female (difference: 4.8pp)
- Present /r/s coded *much* better when speaker is male (difference: 11.3pp)
- These differences mirror the training set's overall /r/ ~ gender correlation



This classifier **fails to satisfy fairness criteria**, likely due to **overlearning** some measures that correlate with gender.

RQ3: Unfairness mitigation

I tested 17 strategies in 4 categories suggested by AI fairness literature [e.g., 10]:

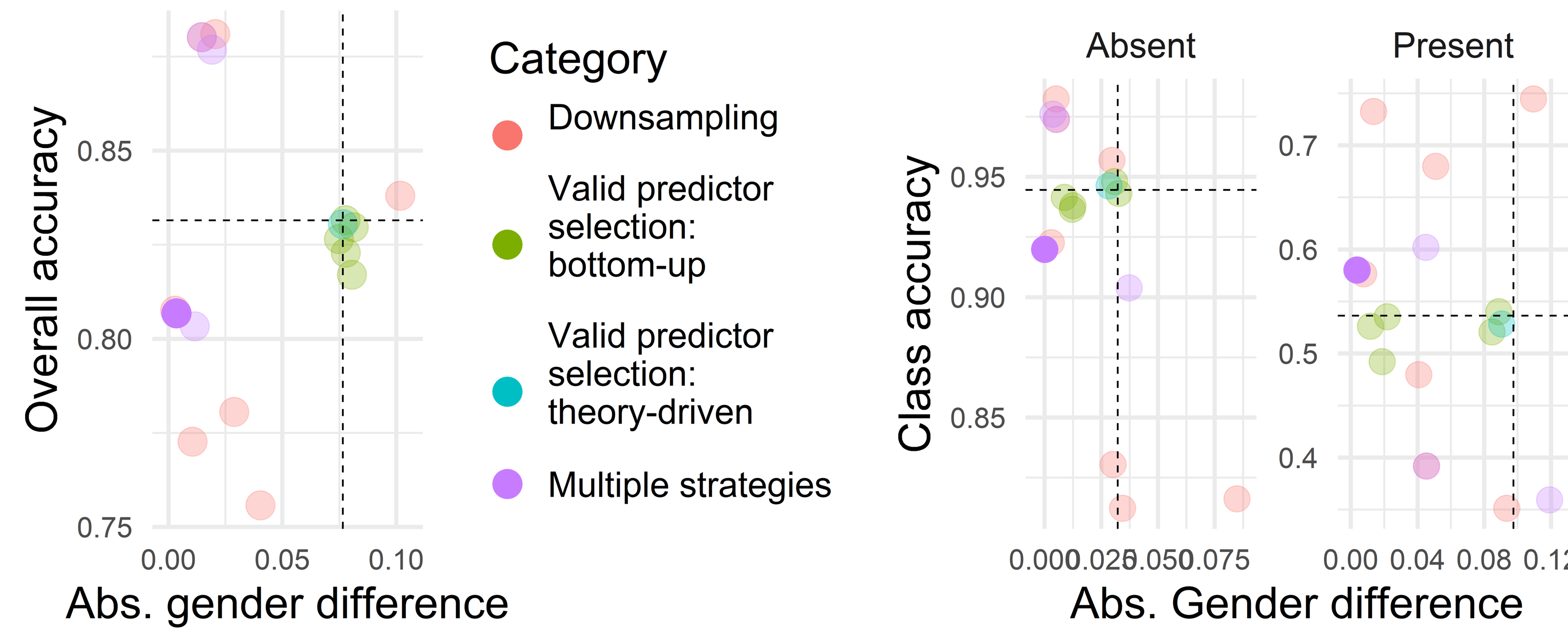
(1) Downsampling (7 versions tested)
- Randomly select data to remove, to correct for imbalances in training data

(2) Valid predictor selection: bottom-up (5 versions tested)
- Remove acoustic measures associated with gender in the model

(3) Valid predictor selection: theory-driven (1 version tested)
- Remove acoustic measures known to be associated with gender (i.e., F0)

(4) Combinations of other strategies (4 versions tested)

Results: Unfairness mitigation strategies



- Numerous strategies improved on baseline's OAE & CAE
  - Of the 17 strategies tested, 10 produced nonsignificant differences between women's & men's overall accuracy
- Fairness maximized by a combination strategy: **removing F0 measures + downsampling** (removing female Absent to get equal /r/ base rates by gender)
  - Compared to baseline (RQ2), this model performed worse for Absent accuracy, but *better* for Present
  - Gender unfairness due to /r/ base rates, not the size of gender training sets

Discussion

- Sociolinguistic auto-coding is not immune to AI unfairness
  - Here, unfairness caused by overlearning speaker gender from acoustics & uneven base rates
- SLAC's characteristics are distinct from other AI fairness use cases
  - Because we hypothesize inter-group differences, statistical parity is undesirable in an auto-coder
- Mitigating cross-group unfairness in SLAC is possible, albeit at the expense of overall performance
  - Worth the tradeoff if group is pertinent to research questions that auto-coded data will be used for

References

This research was supported in part by the Univ of Pittsburgh Center for Research Computing through the resources provided.

[1] Kendall, Tyler et al. 2021. Considering performance in the automated and manual coding of sociolinguistic variables: Lessons from variable (ING). *Frontiers in Artificial Intelligence* 4. doi: 10.3389/frai.2021.648543.

[2] McLarty, Jason, Taylor Jones, and Christopher Hall. 2019. Corpus-based sociophonetic approaches to postvocalic r-lessness in African American Language. *American Speech* 94. doi: 10.1215/00031283-7362239.

[3] Villarreal, Dan et al. 2020. From categories to gradience: Auto-coding sociophonetic variation with random forests. *Laboratory Phonology* 11: 1–31. doi: 10.5334/labphon.216.

[4] Koenecke, Allison et al. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*: 201915768. doi: 10.1073/pnas.1915768117.

[5] Berk, Richard et al. 2021. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* 50: 3–44. doi: 10.1177/0049124118782533.

[6] Villarreal, Dan et al. 2019. How to train your classifier. doi: https://nzilbb.github.io/How-to-Train-Your-Classifier/How_to_Train_Your_Classifier.html.

[7] R Core Team. 2021. *R: A language and environment for statistical computing.*

[8] Kuhn, Max. 2021. *Caret: Classification and regression training.*

[9] Wright, Marvin N., Stefan Wager, and Philipp Probst. 2021. *Ranger: A fast implementation of random forests.*

[10] Corbett-Davies, Sam et al. 2017. Algorithmic decision making and the cost of fairness. doi: 10.1145/3097983.3098095.