Overlearning speaker gender in sociolinguistic auto-coding: Metrics and remedies

Dan Villarreal

March 18, 2022

Penn Linguistics Conference (Special Panel: Sociophonetics and Human-computer Interactions)

Overview

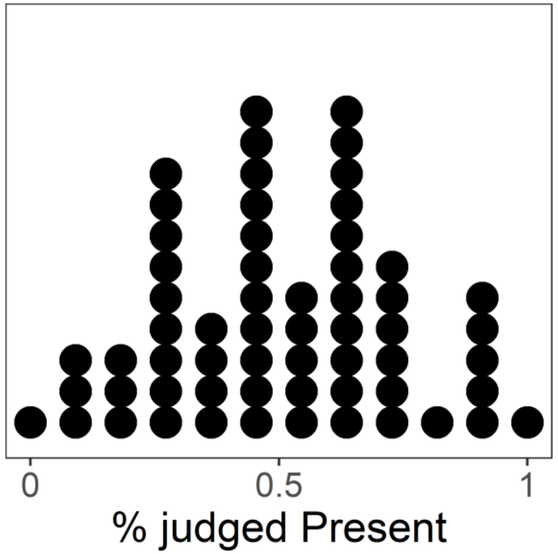
- Today I'll discuss **sociolinguistic auto-coding (SLAC)**, a computational method for classifying variable linguistic data based on acoustic features (Kendall et al. 2021; McLarty et al. 2019; Villarreal et al. 2020)
- I consider SLAC through the lens of **AI fairness** in three steps:
  1. Defining fairness in the context of SLAC
  2. Assessing fairness in a SLAC use case

     Spoiler alert: SLAC is unfair!
  3. Mitigating unfairness via several strategies

     Spoiler alert: SLAC unfairness can be mitigated—but at the cost of auto-coding accuracy

Sociolinguistic auto-coding in a nutshell

- If we want to do a variationist analysis of a categorical variable, we first need to code the variable: label tokens with variants
- But coding is hard
  - Time-consuming, tedious labor
  - /r/ notoriously hard to code, with inter-coder reliability % in the low 80s (Fosler-Lussier et al. 2007; Lawson et al. 2014; Pitt et al. 2005; Yaeger-Dror et al. 2009)
    - Right: 11 phonetically trained listeners judged 60 /r/ stimuli as Present/Absent. Most stimuli received little agreement (Villarreal et al. 2020)
- The solution? Make computers do the work!
  - Like other pinch-points in the sociolinguistic research workflow: phonetic alignment (e.g., McAuliffe et al. 2017), vowel measurement (e.g., Barreda 2021), transcription (e.g., Wassink et al. 2018)



Sociolinguistic auto-coding in a nutshell

- Enter **sociolinguistic auto-coding (SLAC)**!
- Villarreal et al. (2020): https://is.gd/djv012
  - 4,689 hand-coded /r/ tokens, 72.2% Absent (more on the speech community soon)
  - 180 acoustic measures (many speaker-normalized): formants (at numerous timepoints), pitch, amplitude, timing
  - Implemented using random forests in R w/ packages `caret` & `ranger`
  - Performance assessed using cross-validation (training data ≠ test data)
- Auto-coder performed well enough:
  - Accuracy: 84.5% (compares favorably to human inter-coder reliability)
    - That is, the auto-coder identified the correct variant for 84.5% of tokens
  - 'True positive' rate for minority Present class (62.2%) poor compared to majority Absent class (93.1%)
    - That is, the auto-coder identified only 62% of Present tokens as Present!

But…is it **fair**?

…and if it wasn’t, **how would we know**?

AI fairness

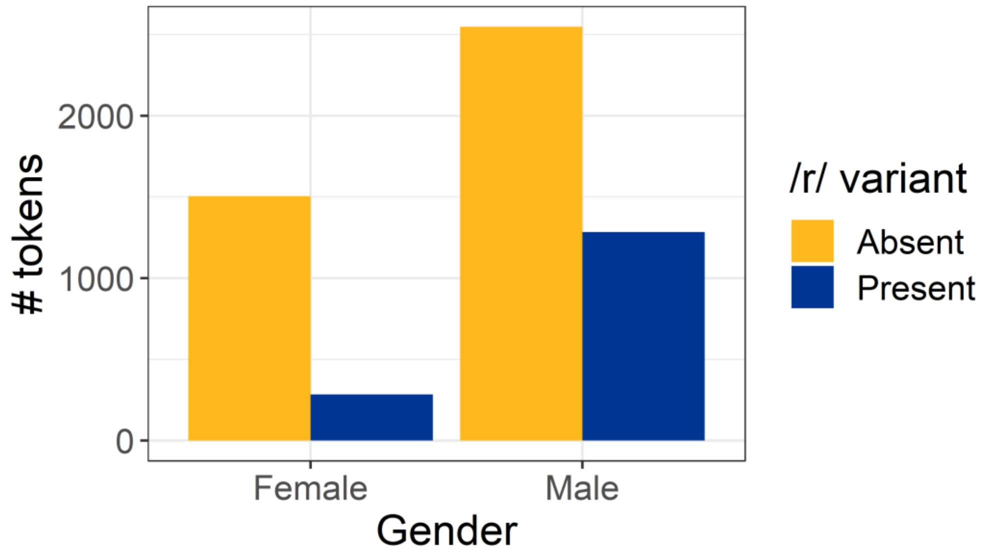
- Angwin et al. (2016) raised concerns that COMPAS, a proprietary algorithm assessing the risk of a pretrial defendant, inadvertently uses defendants' race as a decision criterion
  - Race is *not* an explicit part of the training data for COMPAS
  - Yet COMPAS **overlearns** race by implicitly recovering racial identification from questions such as those about parents' criminal record and peers' drug use
  - As a result, lower-risk Black defendants are erroneously classified as higher-risk
- In theory, SLAC could also be prone to predictive bias, if it makes predictions based *not* on legitimate cues to class membership, but instead inadvertently on group membership

Fairness metrics

- Theoretical research on AI fairness has found that there are multiple **fairness metrics**—and they're (almost always) mutually incompatible (e.g., Berk et al. 2018; Corbett-Davies et al. 2017; Kleinberg et al. 2017)
  - Mutually compatible *only* when base rates are identical (Berk et al. 2018)
- As a result, there is no single ideal fairness metric for all AI applications
- I argue an optimally fair /r/ auto-coder should minimize these three:
  - **Overall accuracy difference**: How much better/worse does the auto-coder work for *all* tokens (regardless of variant) from women vs. men?
  - **Absent class accuracy difference**: How much better/worse does the auto-coder work for *Absent* tokens from women vs. men?
  - **Present class accuracy difference**: the same, but for *Present* tokens
- Other metrics like statistical parity (equal % of predicted Absent/Present for women vs. men) don't make sense for SLAC

Fairness in Villarreal et al. (2020) auto-coder

- Data from Southland New Zealand English, NZ's only regional accent
  - Variably rhotic, unlike non-rhotic General NZE
- In the NZ popular imagination, rhoticity is linked with rugged, rural masculinity considered iconic of Southland (Villarreal et al. 2021)
- Indeed, in the training set, men are significantly more rhotic than women
  - Men's tokens also outnumber women's 2-to-1
- If unfairness is simply a matter of representation in the data, then men should be coded better than women
- Assessed fairness by breaking down performance by speaker gender



Fairness in Villarreal et al. (2020) auto-coder

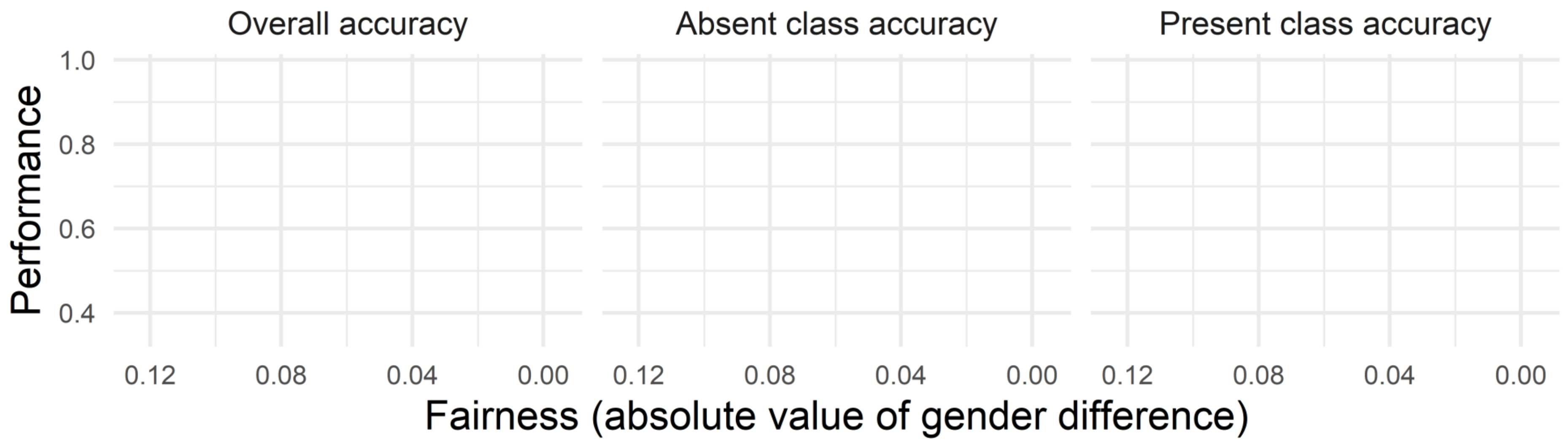
- All 3 metrics: **unfair** 😱
  - Statistically significant differences
- For overall accuracy, women > men
  - Size of training set doesn't guarantee good auto-coding performance
- Women coded better for Absent, men coded better for Present
  - Mirrors the training set's overall /r/ ~ gender pattern
- As a result, this classifier's failure to satisfy SLAC fairness criteria is likely due to **overlearning** some features that correlate with gender

	Female	Male	F − M
OvAcc	89.1%	82.2%	+6.9pp

Unfairness mitigation: Methods

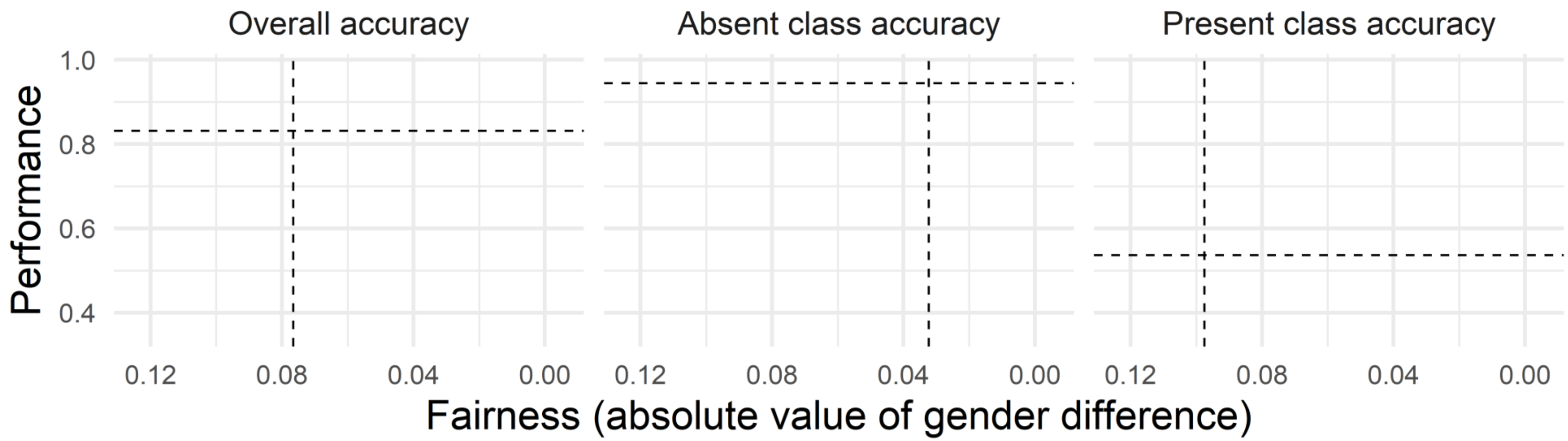
- Various **unfairness mitigation strategies** (UMSs) have been suggested, many of which amount to “systematically hide information from the model that might be contributing to unfairness”
- I tested 4 types of UMS, with 17 different implementations:
  - **Downsampling** (7 implementations): Randomly select observations to remove, to correct for imbalances in training data
  - **Valid predictor selection: empirical** (5 implementations): Remove acoustic measures empirically associated with gender
  - **Valid predictor selection: theoretical** (1 implementation): Remove acoustic measures known to be associated with gender (i.e., pitch)
  - **Combinations of above strategies** (4 implementations)

Unfairness mitigation: Results



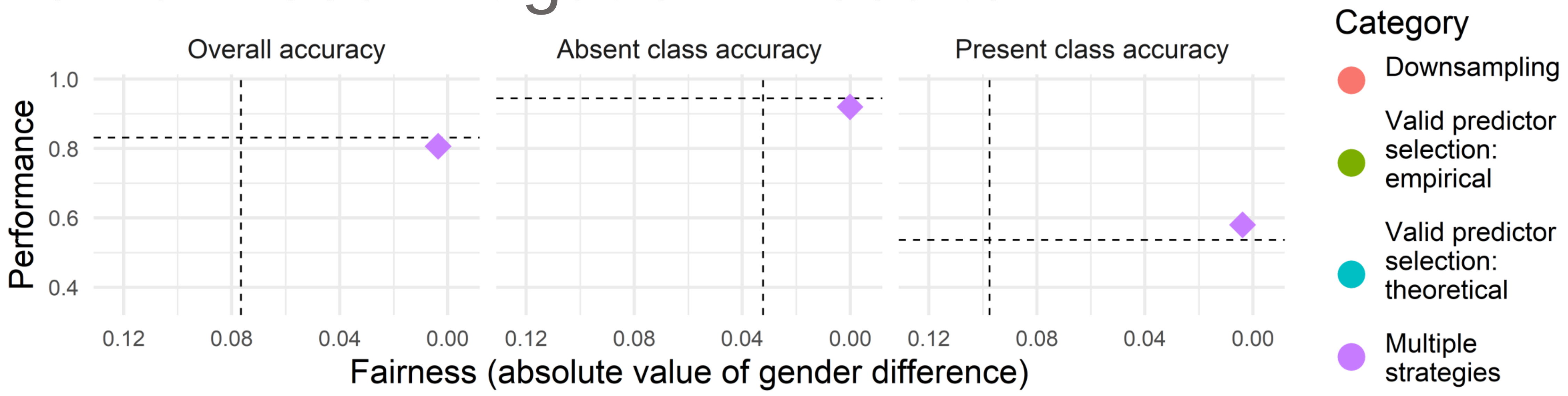
- For each sub-plot, top right = best performance **and** most fair

Unfairness mitigation: Results



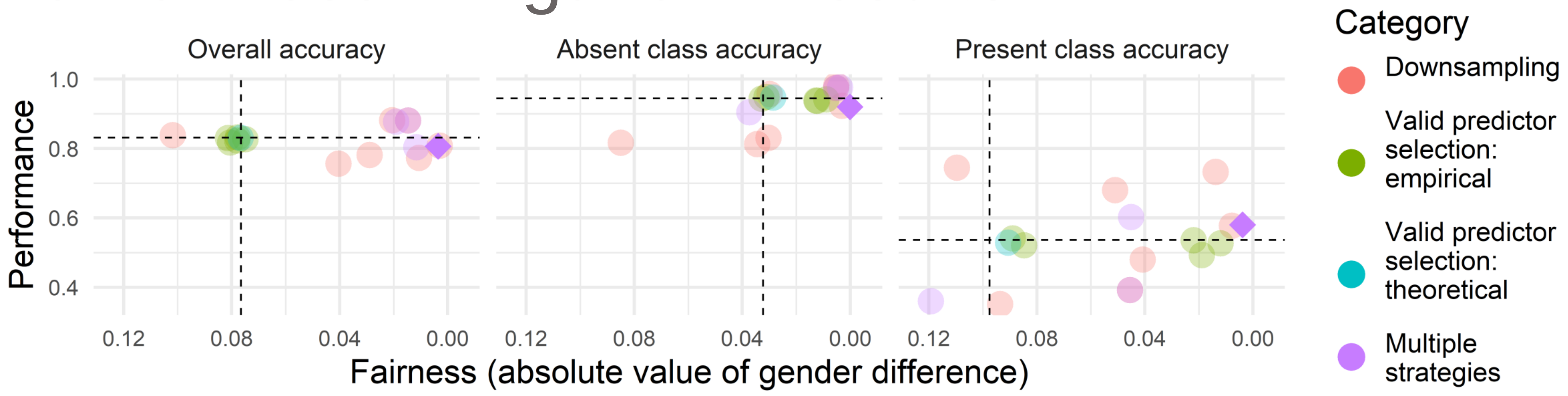
- For each sub-plot, top right = best performance **and** most fair
- Dotted line = baseline classifier from Villarreal et al. (2020)

Unfairness mitigation: Results



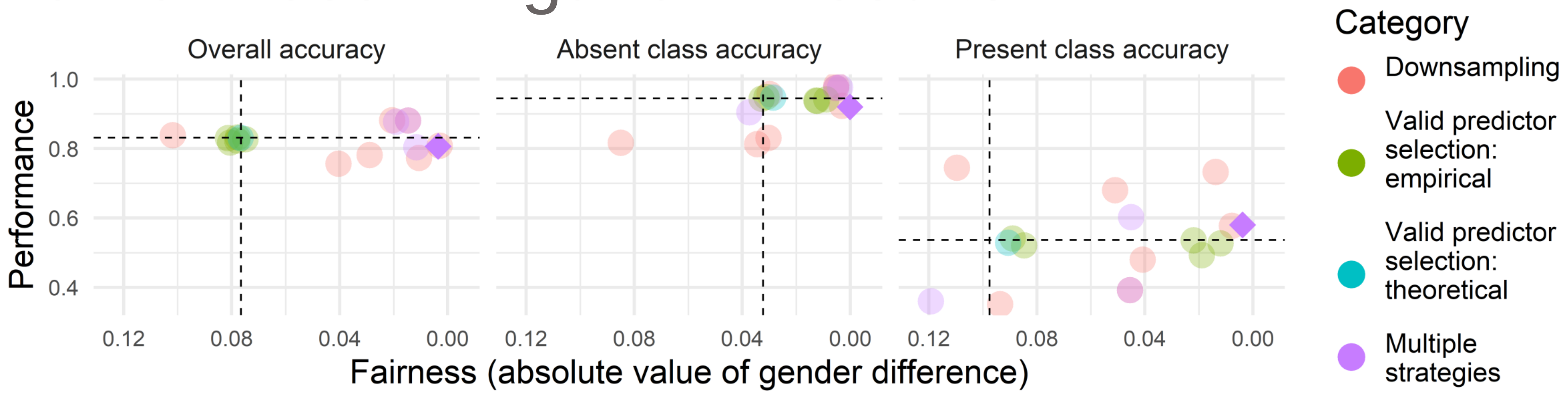
- For each sub-plot, top right = best performance **and** most fair
- Dotted line = baseline classifier from Villarreal et al. (2020)
- Diamond = UMS with maximal fairness

Unfairness mitigation: Results



- For each sub-plot, top right = best performance **and** most fair
- Dotted line = baseline classifier from Villarreal et al. (2020)
- Diamond = UMS with maximal fairness
- Dots = Sub-optimal UMSs
- So what does all this mean?

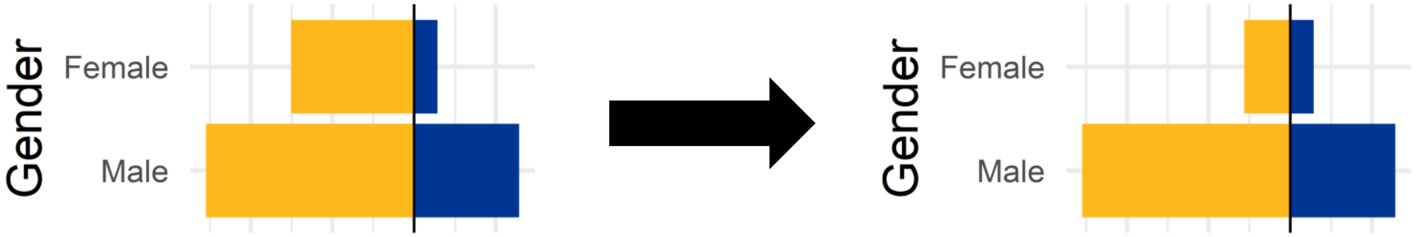
Unfairness mitigation: Results



- Numerous strategies improved fairness
  - 10 of 17 UMSs yielded no significant differences in overall accuracy
- Some strategies failed to improve fairness, or even exacerbated bias
- Many (but not all) strategies that improved fairness worsened performance
  - This includes the UMS with maximal fairness

Unfairness mitigation: Results

- Optimal UMS combined **downsampling** (decreasing female Absent to yield equal /r/ base rates by gender) **+ removing 4 pitch features**



- Unfairness shrank to near-zero
- This came at the expense of classes where women/men had previously been over-represented
  - For example, men's Present accuracy dropped 6.3pp from 64.4%

	Female	Male	F − M
OvAcc	80.4%	80.7%	**-0.4pp**
ClassAcc Absent	92.0%	92.0%	**+0.001pp**
ClassAcc Present	57.1%	58.1%	**-0.4pp**

Sociolinguistic auto-coding and AI fairness

- Like other AI applications, SLAC is indeed prone to predictive bias
  - In this case, gender unfairness is caused by overlearning an association between speaker gender and acoustic features in the feature set, and
- SLAC represents a unique use case for understanding AI fairness
  - Statistical parity—arguably the most important fairness criterion in use cases like pretrial risk detection—is wholly inappropriate for SLAC
- Mitigating cross-group unfairness in SLAC is possible, albeit at the expense of overall performance (accuracy)
  - When we used auto-coded data to compare gender differences in rhoticity in Southland, we decided the performance loss was worth it (Villarreal et al. 2021)
- **Algorithms are never, *never* neutral**. They're the result of choices by human designers—so they reflect our priorities, biases, and mistakes

Thanks! Questions?

Thanks to Lynn Clark, Jen Hay, Kevin Watson, Vica Papp, James Brand, Jacq Jones, Marie Fournier, Simon Todd, Donald Derrick, Jonathan Dunn, and Christina Calvillo
This research was supported in part by the University of Pittsburgh Center for Research Computing through the resources provided.

Dan Villarreal
d.vill@pitt.edu
github.com/djvill