Dan Villarreal*

Sociolinguistic auto-coding has fairness problems too: measuring and mitigating bias

https://doi.org/10.1515/lingvan-2022-0114
Received September 12, 2022; accepted June 28, 2023; published online March 12, 2024

Abstract: Sociolinguistics researchers can use sociolinguistic auto-coding (SLAC) to predict humans' hand-codes of sociolinguistic data. While auto-coding promises opportunities for greater efficiency, like other computational methods there are inherent concerns about this method's *fairness* – whether it generates equally valid predictions for different speaker groups. Unfairness would be problematic for sociolinguistic work given the central importance of correlating speaker groups to differences in variable usage. The current study examines SLAC fairness through the lens of gender fairness in auto-coding Southland New Zealand English non-prevocalic /r/. First, given that there are multiple, mutually incompatible definitions of machine learning fairness, I argue that fairness for SLAC is best captured by two definitions (overall accuracy equality and class accuracy equality) corresponding to three fairness metrics. Second, I empirically assess the extent to which SLAC is prone to unfairness; I find that a specific auto-coder described in previous literature performed poorly on all three fairness metrics. Third, to remedy these imbalances, I tested unfairness mitigation strategies on the same data; I find several strategies that reduced unfairness to virtually zero. I close by discussing what SLAC fairness means not just for auto-coding, but more broadly for how we conceptualize variation as an object of study.

Keywords: sociolinguistic auto-coding; computational and corpus methods; language variation and change; machine learning; bias

1 Introduction

As research in language variation and change involves ever-larger corpora, researchers have increasingly turned to computational techniques to relieve pinch-points in the methodological workflow (e.g., Barreda 2021; McAuliffe et al. 2017; Wassink et al. 2018). Before sociolinguistic variation can be analyzed, for example, researchers must perform tedious and time-consuming *coding:* assigning variants to tokens. Thus, several teams of researchers have in recent years independently explored *sociolinguistic auto-coding (SLAC)* for phonological variables (Kendall et al. 2021; McLarty et al. 2019; Villarreal et al. 2020). In a typical phonological[1] SLAC use case, humans hand-code a fraction of the available tokens, rather than the entire data set; these hand-coded tokens (and their acoustic characteristics) comprise the *training set* from which a machine learning (ML) model attempts to discern the acoustic patterns characteristic of different variants. This model (also known as an *auto-coder*) then applies these patterns to a *test set* of uncoded tokens to predict how humans would have coded them, essentially replicating human coding with substantial time savings. These early investigations have found auto-coding performance comparable to human inter-coder reliability.

Research on *ML fairness* has found that predictive algorithms can reproduce intergroup biases in the data they are trained on (e.g., Field et al. 2021; Koenecke et al. 2020), with concrete costs to potential users

1 While this paper focuses on phonological variation, questions of fairness are equally relevant for (semi-)automated processes on all levels of language: language identification (Blodgett and O'Connor 2017), word sense disambiguation (Austen 2017), and other natural language processing tasks (Dunn 2022). Thanks to Jack Grieve for bringing up this point.

***Corresponding author: Dan Villarreal**, Department of Linguistics, University of Pittsburgh, 2816 Cathedral of Learning, Pittsburgh, PA 15260, USA, E-mail: d.vill@pitt.edu. https://orcid.org/0000-0002-6070-1138

(Mengesha et al. 2021). For example, journalists have raised concerns that an algorithm used to assess the risk of a pretrial defendant, COMPAS, inadvertently uses defendants' race as a decision criterion (Angwin et al. 2016). COMPAS erroneously classifies lower-risk Black defendants as higher-risk, suggesting that its risk classifications are based (partially, at least) on the defendant's race. COMPAS's training set does not explicitly include race, but it does include items from which COMPAS can implicitly recover racial identification (e.g., parents' criminal record, peers' drug use). This predictive bias thus stems from a phenomenon I call "overlearning": when an algorithm's predictions are inadvertently based at least in part on group membership.[2] As a rule, ML fairness is generally an afterthought in an algorithm's development (e.g., Bender et al. 2021; Field et al. 2021); several US states had been using COMPAS risk scores for years before its fairness issues came to light (Angwin et al. 2016). Making matters more complicated, there is no universally applicable way to define ML fairness (e.g., Berk et al. 2021; Corbett-Davies et al. 2017; Kleinberg et al. 2017).

As an ML application, SLAC is potentially subject to biased predictions as a result of overlearning group-level characteristics rather than (or in addition to) legitimate acoustic markers of different variants. Whereas the cost of COMPAS's biased predictions is wrongful detention, the cost of biased predictions in SLAC would be erroneous empirical observations. Given the central importance in sociolinguistics of correlating speaker groups to differences in variable usage, an auto-coder that under- or over-represented the strength of a speaker group constraint would be highly problematic. Fortunately, SLAC is in its infancy; unlike COMPAS and other biased algorithms, it is not too late to reverse the typical "unleash the algorithm first, ask questions later" pattern. The present study thus seeks to answer the following research questions (RQs):
(1) How should (un)fairness be defined and measured for SLAC?
(2) Is SLAC prone to unfairness? If so, then …
(3) How well do unfairness mitigation strategies work for SLAC?

I investigate these research questions with speaker gender as the group-level characteristic, and English non-prevocalic /r/ as the variable to be auto-coded. Both gender and /r/ are good "test cases" for SLAC fairness. Gender fairness is relevant since gender commonly influences variation (e.g., Cheshire 2004; Labov 1990, 2001), including /r/ (e.g., Bartlett 2002; Nagy and Irwin 2010; Villarreal et al. 2021). /r/ is a prototypical variable for SLAC since it is acoustically complex (Heselwood 2009; Lawson et al. 2014, 2018; Stuart-Smith 2007; Stuart-Smith et al. 2014; Zhou et al. 2008), difficult to code reliably (Fosler-Lussier et al. 2007; Hall-Lew and Fix 2012; Irwin 1970; Lawson et al. 2014; Pitt et al. 2005; Yaeger-Dror et al. 2009),[3] usually treated as a binary with present and absent (also known as rhotic and non-rhotic) variants, and variable within and across speech communities (for these two final points, see, e.g., Bartlett 2002; Becker 2009; Gordon et al. 2004; Labov et al. 2006; Nagy and Irwin 2010). Despite these challenges, /r/ auto-coders have demonstrated performance comparable to human inter-coder reliability (Kendall et al. 2021; Villarreal et al. 2020).

This paper investigates these questions from the perspective of researchers who want to use SLAC on data where they suspect gender correlates with variable rhoticity. First, I choose definitions that make sense for SLAC and translate these definitions to quantifiable metrics. Second, I re-analyze Villarreal et al.'s (2019, 2020) /r/ auto-coder, finding that it indeed makes unfair predictions by gender, possibly as a result of overlearning speaker gender. Third, I attempt multiple strategies to mitigate gender unfairness, finding that it is possible to produce a fair auto-coder, albeit at the expense of overall auto-coding performance; one such strategy was used to create a fair /r/ auto-coder in a separate article (Villarreal et al. 2021). I close by discussing what SLAC fairness means not just for auto-coding, but more broadly for how we conceptualize variation as an object of study.

2 ML literature uses the term "bias" more in the statistical sense (e.g., "sampling bias") than the sociological/sociolinguistic sense (e.g., "accent bias"). This paper uses "bias" in the statistical/ML sense. We should acknowledge, however, that the prevalence of (statistical) bias in the real-world use of algorithms is itself a manifestation of structural racism and anti-Blackness (Bender et al. 2021; Markl 2022).
3 Arguably, variables like /r/ that are difficult for humans to code reliably are also those that lack "straightforward acoustic cues" (Kendall et al. 2021: 2) and are thus well suited for ML methods that can learn complex patterns.

1.1 Gender and categoricity

The Villarreal et al. (2020) /r/ auto-coder re-analyzed in this paper used legacy data that only categorizes speakers as female or male (Bartlett 2002). However, this paper's fairness approach is not locked into a binary conception of gender, race, or any other group characteristic. A far harder problem for group fairness is that categoricity itself is somewhat artificial – identifying discrete groups is a methodological convenience that greatly simplifies the diverse, dynamic, and fluid identities that speakers construct through day-to-day performance in interaction (e.g., Eckert and McConnell-Ginet 1992; Holliday 2019; Zimman 2018).[4] That is, it is much easier to measure auto-coding fairness between discrete genders than across the spectrum of gender performances in speech. Grappling with the reality of dynamic identities continues to be at the frontier in computational sociolinguistics research (Charity Hudley et al. 2023; Nguyen et al. 2014). Nevertheless, to the extent that future sociolinguistics research asks questions about discrete genders, it is useful to know whether auto-coding is fair across these groups.

2 RQ1: Defining and measuring fairness for SLAC

Algorithms that attempt to sort observations into two or more categories are known as "classifiers" (e.g., SLAC, spam filtering, cancer detection), with the categories known as "classes" (e.g., present vs. absent /r/, spam vs. not spam, malignant vs. benign; Hastie et al. 2009). Classifier *performance* is assessed by comparing the classifier's predictions to the so-called ground truth. For two-class classifiers, such as /r/ auto-coders, this comparison can be represented in a confusion matrix, such as Table 1; a "true present", for example, is an /r/ token that is present in reality and was correctly auto-coded "present". These concepts are similar to notions like "true positive" and "false positive" (which are common in ML, medical testing, and other domains); but "positive/negative" does not make sense for auto-coding because it implies that only one class ("positive") is the real object of detection, as with classifiers that attempt to detect spam email or malignant tumors. Arguably, in auto-coding "no class is more important for purposes of detection" (Villarreal et al. 2020: 11).

The confusion matrix is the building block for calculating performance metrics (quantifiable measurements of a classifier's success at sorting observations into classes). Table 2 displays some common performance metrics. Analysts typically use performance metrics to compare classifiers to one another; for example, Kendall et al. (2021) find that overall accuracy for *-ing* classifiers changes based on the type of training data, and Villarreal et al. (2020) find better overall accuracy for binary /t/ than binary /r/. However, these metrics can also be used to assess fairness, by breaking down performance by group (Berk et al. 2021; Corbett-Davies et al. 2017). Thus, just as we can quantify performance, we can quantify fairness for the purposes of choosing a fair auto-coder.

There are multiple possible definitions of fairness, and if groups have different base rates (see Table 2), it is impossible for a classifier to satisfy all fairness definitions at once (Berk et al. 2021; Corbett-Davies et al. 2017; Kleinberg et al. 2017). Applied to SLAC and gender fairness, an auto-coder cannot satisfy all definitions if women and men have different community-wide rhoticity rates – exactly the sort of external factor sociolinguists

Table 1: A hypothetical /r/ auto-coder's confusion matrix.

		Ground truth	
		Absent	**Present**
SLAC prediction	Absent	True absent (TA)	False absent (FA)
	Present	False present (FP)	True present (TP)

4 Thanks to Zach Jaggers for bringing up this point.

Table 2: Some performance metrics and formulas for a hypothetical /r/ auto-coder.

Performance metric	Overall	Absent	Present
Overall accuracy (OA)	$\frac{TA+TP}{TA+FA+FP+TP}$		
Class accuracy (CA)		$\frac{TA}{TA+FP}$	$\frac{TP}{TP+FA}$
Base rate		$\frac{TA+FP}{TA+FA+FP+TP}$	$\frac{TP+FA}{TA+FA+FP+TP}$
Predicted prevalence		$\frac{TA+FA}{TA+FA+FP+TP}$	$\frac{TP+FP}{TA+FA+FP+TP}$
Predictive value		$\frac{TA}{TA+FA}$	$\frac{TP}{TP+FP}$
Miscoding ratio	$\frac{FP}{FA}$		

Table 3: Fairness definitions and metrics used in the present study.

Definition	Translation	Metric	Formula
Overall accuracy equality	Women's and men's tokens are coded equally well regardless of whether the token is absent or present	Overall accuracy difference	$OA_F - OA_M$
Class accuracy equality	The auto-coder detects absent tokens equally well for women and men	Absent class accuracy difference	$CA_{Abs,F} - CA_{Abs,M}$
	The auto-coder detects present tokens equally well for women and men	Present class accuracy difference	$CA_{Pres,F} - CA_{Pres,M}$

attempt to detect (This is in contrast to classification use cases like pretrial risk assessment, where it is reasonable to assume *a priori* that potential Black and White defendants are equally risky). As a result, analysts must choose the fairness definition or definitions that make the most sense for them (Kleinberg et al. 2017).

For SLAC, I argue that "fairness" is best captured by two definitions – overall accuracy equality and class accuracy equality – corresponding to three metrics (Table 3). These definitions make sense because of what separates SLAC from other classification problems. Because there is no "positive" class (as discussed above), it makes sense to measure both overall accuracy equality, which "assumes that true negatives are as desirable as true positives" (Berk et al. 2021: 15), and class accuracy equality (for both absent tokens and present tokens). Overall accuracy equality is also useful over and above class accuracy equality because it weights each class by its prevalence in the data, thus giving the clearest picture of the total number of (in)accurately coded tokens. Because we cannot assume equal base rates *a priori,* it makes no sense to pursue (conditional) statistical parity (Berk et al. 2021: 16; Corbett-Davies et al. 2017: 798), which would penalize auto-coders for predicting different rates of rhoticity for women and men. These metrics may not capture "fairness" for all potential SLAC use cases, however, so Section 5 revisits alternative fairness definitions and metrics.

3 RQ2: Assessing fairness for SLAC

Using the fairness definitions and metrics established in the previous section, this section re-analyzes Villarreal et al.'s (2019, 2020) /r/ auto-coder to assess gender fairness.

3.1 Methods

RQ2 and RQ3 were addressed with the same data set and auto-coding implementation, based directly on Villarreal et al.'s (2019, 2020) approach. The data came from Bartlett's (2002) doctoral research on Southland New Zealand

English. Variable rhoticity in Southland has historically set it apart from the rest of New Zealand English both in fact and in folk-linguistic belief; in the New Zealand popular imagination, rhoticity is linked with rugged, rural masculinity considered iconic of Southland (Jackson et al. 2009; Villarreal et al. 2021). Bartlett conducted sociolinguistic interviews and performed all hand-coding (with an absent/present binary). Only a subset of his hand-codes were recovered due to issues with old data formats. Out of 30,777 /r/ tokens in his corpus, ground truth hand-codes were available for only 5,620, and this training set skews male (31.8 % female). The training set overall skews absent (27.9 % present), and in line with folk-linguistic belief, men are more rhotic than women (female: 15.9 %; male: 33.5 %). This particular data set only categorizes speakers as female or male.

Villarreal et al.'s (2019, 2020) auto-coder was re-analyzed for RQ2, whereas new auto-coders were run for RQ3. All auto-coders were run on a set of 180 acoustic measures extracted from each token, including formant frequencies at multiple timepoints, pitch, intensity, and timing; formant frequencies were normalized by speaker and preceding vowel, but no other measures were normalized. Acoustic measures were extracted via Praat (Boersma and Weenink 2022) through a LaBB-CAT corpus interface (Fromont and Hay 2012). Readers can visit GitHub to download the data (https://github.com/nzilbb/Sld-R-Data) and Villarreal et al.'s (2020) auto-coder (https://github.com/nzilbb/How-to-Train-Your-Classifier).

Auto-coders were implemented via the random forest method (e.g., Breiman 2001; Tagliamonte and Baayen 2012) in R using the caret and ranger packages (Kuhn 2022; R Core Team 2022; Wright et al. 2021).[5] While many ML methods can be used for classification (Hastie et al. 2009), random forests are preferable to other ML methods when predictors are collinear (Dormann et al. 2013; Kendall et al. 2021; Matsuki et al. 2016; Strobl and Zeileis 2008), as was highly likely for this set of acoustic measures. For more implementation details (including all measures), see Villarreal et al. (2019, 2020: 7–11).

Finally, to assess fairness, I created a combined data set that had each token's speaker gender, actual class, and predicted class, from which overall accuracy difference and absent/present class accuracy differences were calculated (see Table 3). R code for measuring fairness can be found in the GitHub repository (Villarreal 2023). I hypothesized that if there was unfairness, it would be due to the classifier overlearning speaker gender in rhoticity judgments. Under this hypothesis, some predictor or predictors in the training set would allow the auto-coder to learn that cues of male-hood are associated with the present label to a greater degree than are cues of female-hood; thus, for some men's tokens that are actually absent, the auto-coder would attend not to "legitimate" cues to absent-hood but rather to cues of male-hood, and would thus code those tokens as "present". This hypothesis would also mean that unfairness could be mitigated by finding and removing those "illegitimate" cues to gender from the feature set. Later, my findings for RQ3 will show not only that pitch acts as an "illegitimate" cue to gender rather than rhoticity in this data set, but also that overlearning is not the only source of unfairness (see Section 4).

3.2 Results

Villarreal et al.'s (2019, 2020) auto-coder failed to satisfy any of the chosen fairness definitions, with all three metrics revealing significant accuracy differences between women and men (Table 4). In terms of overall accuracy, women were coded better despite having half as many tokens as men in the training set. This case contrasts with ML unfairness that is caused by groups being inadequately represented in the training set (e.g., Koenecke et al. 2020). Women were not coded better across the board, however; the auto-coder was better at coding absent tokens when they come from women and present tokens from men. In other words, the auto-coder performed better at coding the class for which each gender was better represented in the training set. This pattern suggests support for the overlearning hypothesis, as it appears to associate women with absent to a greater degree than men, and men with present to a greater degree than women.

The practical consequences of this degree of unfairness are alarming, as it could substantially undermine any analyses using auto-coded data (Table 5). We can consider our training set a sample in the sense of inferential

5 The original and new auto-coders were run with R versions 3.5.2 and 4.2.0, caret versions 6.0.84 and 6.0.93, and ranger versions 0.11.1 and 0.14.1, respectively.

Table 4: Gender fairness metrics in Villarreal et al.'s (2019, 2020) /r/ auto-coder. Positive differences indicate better performance for women than men. The χ^2 column reports test of homogeneity comparing female and male columns.

Metric	Female	Male	Difference	χ^2
Overall accuracy	0.8915	0.8222	+0.0693	χ^2 [1] = 37.03, $p < 0.0001$
Absent class accuracy	0.9634	0.9109	+0.0525	χ^2 [1] = 33.18, $p < 0.0001$
Present class accuracy	0.5144	0.6463	−0.1319	χ^2 [1] = 14.06, $p < 0.001$

Table 5: Actual versus predicted gender/rhoticity distribution in training set for Villarreal et al.'s (2019, 2020) /r/ auto-coder training set. Predictions generated by three-repeat twelve-fold cross-validation.

Gender	Class	Actual	Predicted	Under/overprediction
Female	Absent	1,275	1,349	+5.8 %
Female	Present	243	169	−30.5 %
Male	Absent	2,109	2,292	+8.7 %
Male	Present	1,062	879	−17.2 %

statistics, where it is assumed that with respect to the distribution of some variable of interest, the population and a sample (even a representative one) will diverge. While we therefore do not expect that auto-coders will predict the same rate of rhoticity in our training and test sets, we do want auto-coders to make predictions solely on the basis of legitimate indicators of rhoticity. If the auto-coder's predictions were to exaggerate the gender/rhoticity distribution in the training set, this raises the troubling prospect of Type I errors. Hypothetically, this training set could be unluckily selected from a full data set (training + test) in which women and men actually exhibit identical rhoticity; if so, women would be slightly more rhotic than men in the test set. In this case, this unfair auto-coder, having overlearned a coincidental "men favor present" pattern, would predict the same in the test set, causing us to incorrectly claim a gender/rhoticity correlation in the population. Simply put, this would be a very bad outcome.

4 RQ3: Mitigating SLAC unfairness

The unfair predictions in Villarreal et al.'s (2019, 2020) auto-coder potentially compromise SLAC's usefulness in sociolinguistic methodology, as cross-group comparisons are a cornerstone of sociolinguistic research. Fortunately, ML researchers have found numerous ways to mitigate unfairness in classification problems like SLAC (e.g., Corbett-Davies et al. 2017).

4.1 Methods

I investigated several *unfairness mitigation strategies* (UMSs) using the same data and auto-coding setup as RQ2 (see Section 3.1). These UMSs comprised four basic types, totaling 23 implementations:

(1) Downsampling (seven implementations): correct for imbalances in training data by randomly selecting tokens to remove
(2) Valid predictor selection (seven implementations): remove acoustic measures that could inadvertently signal gender
(3) Normalization (one implementation): control for acoustic variability that could inadvertently signal gender
(4) Combinations of other strategies (eight implementations): downsampling plus valid predictor selection or normalization

These categories all respond to possible underlying causes of ML unfairness. Valid predictor selection attempts to preserve “legitimate” cues to absent/present-hood while removing from the predictor set cues that allow the auto-coder to overlearn female/male-hood (e.g., Corbett-Davies et al. 2017). Normalization, which is commonly used in sociophonetics to control for interspeaker variation in vocal tract length (e.g., Barreda 2021), similarly attempts to sort “signal” from “noise”. Downsampling is rooted in the mathematical property that it is impossible for a classifier to satisfy all fairness definitions at once if base rates differ (e.g., Berk et al. 2021). Although rhoticity base rates do differ in the /r/ training set, we can make them equal by removing tokens from one gender or the other. Within these broad strategies are numerous possible implementations; for example, in this data equal base rates can be achieved by downsampling women’s absent tokens or men’s present tokens. Finally, several strategies (added after an initial analysis of the simple strategies) combined downsampling and either valid predictor selection or normalization. Information on each implementation can be found in Appendix A, and full details (including data and R code) can be found in the GitHub repository (Villarreal 2023).

Unlike Villarreal et al.’s (2019, 2020) auto-coder, the auto-coders in this section did not undergo the time-consuming process of optimization for performance (see Villarreal et al. 2019: “Step 4” and “Step 5”). For comparison to these auto-coders, I ran a baseline (no-UMS) auto-coder that was also not optimized for performance, which performed comparably for overall accuracy difference (+0.0766) and slightly worse for class accuracy differences (absent: +0.0323; present: −0.0975) compared to Villarreal et al.’s (2019, 2020) auto-coder.

4.2 Results

The majority of unfairness mitigation strategies were fairer than the no-UMS baseline (Figure 1). In particular, ten UMSs did not yield any significant differences between women’s and men’s performance (with χ^2 [1] ≤ 2.30, $p \geq 0.13$). This fairness did not come for free, however – all strategies that improved fairness worsened performance on at least one metric, and some UMSs improved fairness or performance on one metric but worsened fairness or performance on others. For example, UMS 1.3.2 reduced the overall accuracy difference between women and men but yielded worse present class accuracy regardless of gender (see Figure 2 below). Regardless of

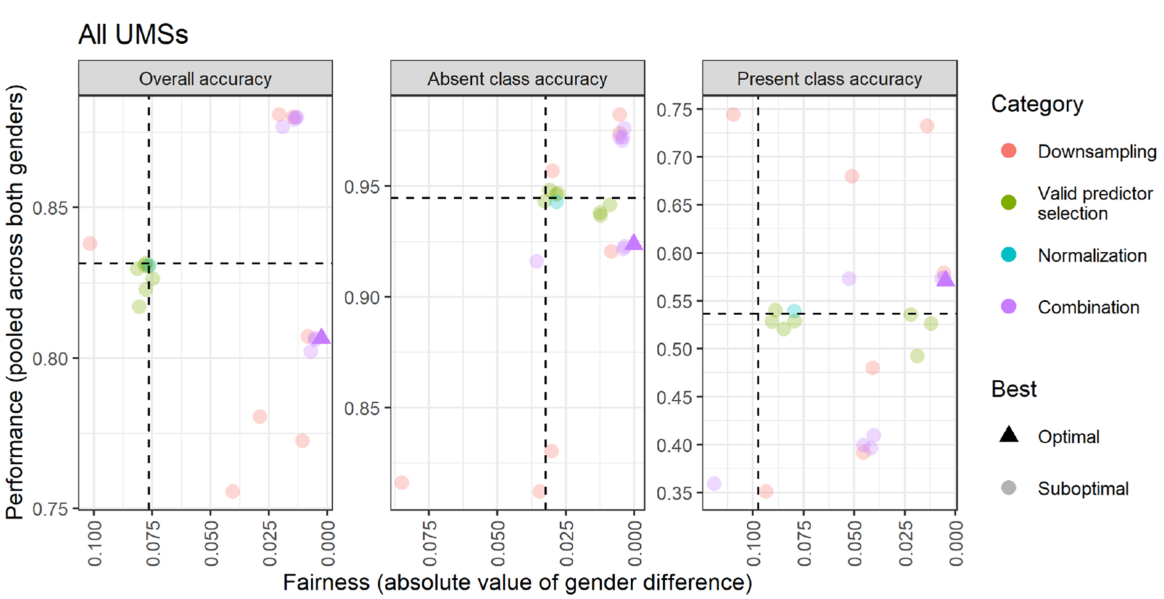

Figure 1: Comparison of the fairness and performance of unfairness mitigation strategies (UMSs). Fairness increases from left to right; performance increases from bottom to top. The dotted line is the no-UMS baseline. “Optimal” denotes the UMS used by Villarreal et al. (2021). The numerical data for this figure can be found in Appendix B.

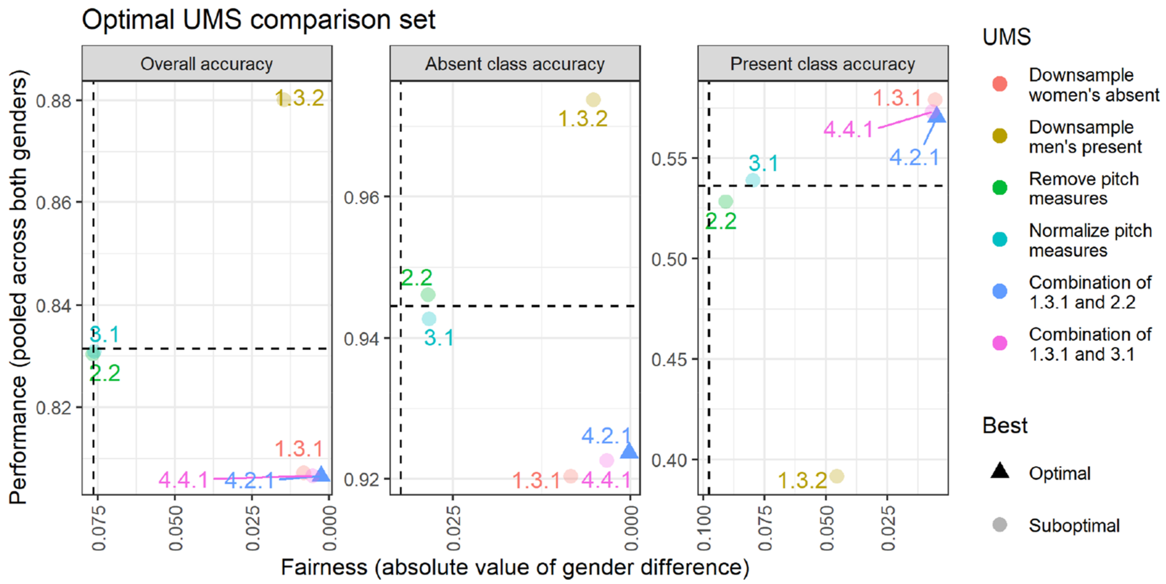

Figure 2: Comparison of the fairness and performance for the optimal unfairness mitigation strategy (UMS) and several similar UMSs. Fairness increases (gender difference decreases) from left to right; performance increases from bottom to top. The dotted line is the no-UMS baseline. UMS labels are as in Appendix A. The numerical data for this figure can be found in Appendix B.

Table 6: Gender fairness metrics for optimal UMS. Values in parentheses indicate the change from the unfair auto-coder results seen in Table 4; positive changes indicate improvement in accuracy (female and male columns) or greater unfairness (difference column). The χ^2 column reports test of homogeneity comparing female and male columns.

Metric	Female	Male	Difference	χ^2
Overall accuracy	0.8044 (−0.0871)	0.8070 (−0.0152)	−0.0026 (−0.0667)	χ^2 [1] = 0.02, p = 0.90
Absent class accuracy	0.9236 (−0.0398)	0.9239 (+0.0130)	−0.0002 (−0.0523)	χ^2 [1] < 0.01, p > 0.99
Present class accuracy	0.5670 (+0.0526)	0.5718 (−0.0745)	−0.0048 (−0.1270)	χ^2 [1] = 0.01, p = 0.93

the strategy, a loss of performance makes intuitive sense, as the UMSs ultimately boil down to “give the auto-coder less information”.

Given this trade-off, the UMS that has the greatest fairness may be undesirable because its performance cost is too high. So how do we choose a UMS for auto-coding our data? One technique for winnowing down the space of options is to find the UMSs for which any other UMS that is better in fairness is worse in performance, or vice versa; in ML, observations that have this property are known as *Pareto-optimal*. Thus, researchers may choose a UMS that is neither the fairest nor the best performing, but for which the fairness–performance trade-off is acceptable.

Villarreal et al. (2021)’s Southland /r/ auto-coder used UMS 4.2.1, which not only is Pareto-optimal in all three facets of Figure 1 but also ranks highest for all three fairness metrics.[6] This UMS is a combination of UMS 1.3.1 (downsampling: removing female absent to achieve equal rhoticity rates by gender) and 2.2 (valid predictor selection: removing F0 measures). UMS 4.2.1 had near-zero unfairness (Table 6). It is important to note that not all auto-coders will necessarily have one UMS that is Pareto-optimal across the board; as a result, SLAC users may face a judgment call in terms of choosing the UMS that is right for their application.

6 The auto-coder used in Villarreal et al. (2021: 252) was additionally optimized for performance (see Section 4.1), so its reported fairness is slightly worse than UMS 4.2.1.

Returning to the hypothesis that overlearning is the underlying cause of unfairness, two pieces of evidence support this hypothesis. First, the auto-coder's class accuracies by gender (absent higher for women, present higher for men) mirror the gender/rhoticity distribution in the training set (men are more rhotic than women). This suggests that, in some cases, the auto-coder attends to acoustic measures that cue gender rather than "legitimate" cues of rhoticity. Second, the unfairness mitigation analysis found acoustic measures (pitch) that, when removed, substantially reduced unfairness.

However, a small comparison set of UMSs tells a different story (Figure 2). If overlearning were the primary cause of unfairness, then removing pitch measures (UMS 2.2) or normalizing pitch by speaker (UMS 3.1) would substantially improve fairness. Both only resulted in modest fairness gains. In other words, overlearning appears to be a cause of unfairness, but not (contra my original hypothesis) the primary cause of unfairness.

5 Discussion

Sociolinguistic auto-coding was developed under the premise that a methodological challenge could be met with a technological solution. Early work in phonological SLAC has demonstrated both its promise for meeting this methodological challenge and that questions remain to be answered (Kendall et al. 2021; McLarty et al. 2019; Villarreal et al. 2020). In short, it is crucial to frame SLAC, COMPAS (Angwin et al. 2016), or any other algorithm not as a "magic bullet"; engaging in uncritical "tech solutionism" prevents us from clearly seeing algorithms' limitations (Bender and Grissom 2024). So too should we avoid the trap of seeing bias successfully mitigated in a single SLAC use case and think "we solved SLAC bias!".[7] This paper used a SLAC use case that I argue is fairly typical: researchers want to deploy an auto-coder of /r/ on large-scale corpus data in a community where they suspect gender will be part of the analysis. But a single use case cannot cover all possible outcomes. Of course, unfair predictions would be problematic anytime speaker groups are in play – since unfair predictions distort correlations between speaker groups and variable usage – but different circumstances may yield different outcomes or even necessitate different fairness definitions. I discuss these different circumstances in reverse order of the research questions.

First, for Southland /r/, unfairness was partially caused by the auto-coder overlearning a group characteristic based on several acoustic measures in the feature set. Other SLAC cases may exhibit fairness metrics that suggest overlearning, but without an acoustic measure (or other feature) that is readily identifiable as the locus of overlearning; in these cases, mitigating unfairness could be more challenging. In other cases, unfairness may not be caused by overlearning at all (e.g., if groups have equal base rates), but inadequate representation in the training set (e.g., Koenecke et al. 2020), necessitating other unfairness mitigation strategies. In still other cases, the UMS that maximizes fairness may negatively impact performance so much that it is untenable to deploy on the test set.

Second, other SLAC use cases may also differ with respect to the degree of unfairness; there is no guarantee that *all* auto-coders will be unfair. Just as Villarreal et al. (2020) find better overall accuracy for binary /t/ than /r/, some variables are easier to code than others (both by hand and by auto-coder); conceivably, these variables may be less prone to unfairness. Likewise, auto-coders that achieve balanced class accuracies (unlike /r/) may be less prone to unfairness. On the other hand, whereas the unfair /r/ auto-coder overpredicted the majority class for both groups, other auto-coders could overpredict different classes for different groups.

Third, I argued that the most appropriate fairness definitions for SLAC were overall accuracy equality and class accuracy equality, but future SLAC use cases may demand alternative fairness definitions.

7 Thanks to Emily Grabowski for proposing this framing.

For some variables in some communities, researchers might discard the "no positive class" assumption that arguably sets SLAC apart from classification problems like spam detection; for example, it may be more important to detect a variant that is incipient or dying out than its better-established alternative(s). Not all categorical variables are binary. In addition, even for traditionally categorical sociolinguistic variables like /r/, sociolinguistic reality may be better represented by continuous acoustic variation (e.g., not absent/present but degree of rhoticity), a view supported by a growing body of sociophonetic research (Duncan 2021; Holliday and Villarreal 2020; Podesva 2007; Purse 2019). When auto-coders are used to generate probabilistic rather than categorical predictions of variant-hood (McLarty et al. 2019; Villarreal et al. 2020), researchers should use fairness definitions that account for probabilistic predictions, such as calibration (Kleinberg et al. 2017). Finally, extending this approach beyond a binary group characteristic would require rethinking how the fairness definitions are translated to quantifiable metrics (e.g., taking the standard deviation of overall accuracies rather than the difference).

Given this discussion, it seems prudent to revisit the initial framing of SLAC as essentially replicating human coding. Instead, these findings support the idea that auto-coders "should not seek simply to replace human analysts but rather that they reflect alternative approaches to coding that have advantages and appropriateness for some applications and disadvantages and inappropriateness for others" (Kendall et al. 2021: 15). In other words, auto-coding (and human coding) is not "searching for one right answer" – different analyses may necessitate different approaches to auto-coding. In the case of Southland /r/, there was previous sociolinguistic and sociological evidence to hypothesize the importance of speaker gender (Bartlett 2002; Jackson et al. 2009), so achieving gender fairness was well worth the performance loss compared to the unfair auto-coder. If a researcher were to perform a future analysis of the same data that only considered men, however, they should use auto-codes that maximize performance. A different future analysis that attempted to cluster speakers based on rhoticity patterns (e.g., Brand et al. 2021) would need to ensure fairness across speakers. In other words, the "correct" code for any given token in the test set may change based on the needs of the analysis. As a result, SLAC is potentially appropriate only for corpus sociolinguistic approaches, rather than micro-sociolinguistic analyses. This may be an unsettling idea for researchers who are used to hand-codes representing the "ground truth", but as corpus sociolinguistics matures, we will need to recognize the inherent uncertainty that "big data" implies. These observations add a new dimension to Villarreal et al.'s (2019) observation that auto-coding is "not a one-size-fits-all process".

These differences notwithstanding, the present study represents two "proofs of concept" – SLAC *is* prone to predictive bias, and SLAC bias *can* be mitigated. To be clear, I am not suggesting that SLAC is too inherently flawed to use; the findings of RQ3 clearly show otherwise, and this bias-mitigation approach to auto-coding has already been used in sociolinguistic research (Villarreal et al. 2021). Rather, I argue that users of computational methodologies like SLAC deserve a "warts and all" awareness of their drawbacks before they have the chance to unleash harm. Algorithms are never neutral – they are the result of choices by human designers, so they reflect our priorities, biases, and mistakes.

Acknowledgments: I would like to thank Chris Bartlett, the Southland Oral History Project (Invercargill City Libraries and Archives), and the speakers for sharing their data and their voices. Thanks are also due to Lynn Clark, Jen Hay, Kevin Watson, and the New Zealand Institute of Language, Brain and Behaviour for supporting this research. Valuable feedback was provided by two anonymous reviewers, James Stanford, and audiences at NWAV 49, the Penn Linguistics Conference, Pitt Computer Science, and the Michigan State SocioLab. Other resources were provided by a Royal Society of New Zealand Marsden Research Grant (16-UOC-058) and the University of Pittsburgh Center for Research Computing (specifically, the H2P cluster supported by NSF award number OAC-2117681). Any errors are mine entirely.

Appendix A: Unfairness mitigation strategy descriptions

The following table provides brief descriptions for the 23 unfairness mitigation strategies (UMSs) tested in this study. UMSs are numbered with the first digit for the category, the second digit for the strategy, and a possible third digit for different implementations of the same strategy. UMS0.0 is the baseline (no-UMS) auto-coder. In the Description column, "Rpresent" refers to the token's class (present/absent).

Category	UMS	Description
Baseline	0.0	Baseline auto-coder of Rpresent, with all data and predictors
Downsampling	1.1	Downsample men to equalize token counts by gender
	1.2	Downsample absent to equalize token counts by Rpresent
	1.3.1	Downsample women's absent to equalize Rpresent base rates by gender
	1.3.2	Downsample men's present to equalize Rpresent base rates by gender
	1.4	Downsample men's data to equalize (a) token counts by gender and (b) Rpresent base rates by gender
	1.5	Downsample absent data to equalize (a) token counts by Rpresent and (b) gender base rates by Rpresent
	1.6	Downsample gender × Rpresent to equalize token counts by gender × Rpresent
Valid predictor selection	2.1.1	Empirical predictor selection, removing most influential predictors in classifier of gender (cutoff: top 10 %)
	2.1.2	Empirical predictor selection, removing most influential predictors in classifier of gender (cutoff: top 20 %)
	2.1.3	Empirical predictor selection, removing most influential predictors in classifier of gender (cutoff: top 50 %)
	2.1.4	Empirical predictor selection, without measures with differential importance in separate-gender auto-coders of Rpresent (difference in rank places: at least p/2)
	2.1.5	Empirical predictor selection, without measures with differential importance in separate-gender auto-coders of Rpresent (difference in rank places: at least p/3)
	2.2	Theoretical predictor selection, removing all F0 measures
	2.3	Empirical and theoretical predictor selection, removing only F0 measures that correlate with gender
Normalization	3.1	Normalize F0 by speaker
Combination	4.1.1	Combination of 2.1.1 and 1.3.1
	4.1.2	Combination of 2.1.1 and 1.3.2
	4.2.1	Combination of 2.2 and 1.3.1
	4.2.2	Combination of 2.2 and 1.3.2
	4.3.1	Combination of 2.3 and 1.3.1
	4.3.2	Combination of 2.3 and 1.3.2
	4.4.1	Combination of 3.1 and 1.3.1
	4.4.2	Combination of 3.1 and 1.3.2

Appendix B: Unfairness mitigation strategy fairness and performance

The following table presents the data from Figures 1 and 2. UMS0.0 (the no-UMS baseline auto-coder) is the dotted line in those figures.

UMS	Overall accuracy	Overall accuracy difference	Absent class accuracy	Absent class accuracy difference	Present class accuracy	Present class accuracy difference
0.0	0.8315	+0.0766	0.9445	+0.0323	0.5363	−0.0975
1.1	0.8380	+0.1017	0.9568	+0.0298	0.4795	−0.0408
1.2	0.7806	+0.0289	0.8161	+0.0849	0.7443	−0.1096
1.3.1	0.8072	−0.0082	0.9204	−0.0084	0.5791	−0.0055
1.3.2	0.8801	+0.0146	0.9738	+0.0052	0.3917	+0.0455
1.4	0.8809	+0.0206	0.9823	+0.0051	0.3510	+0.0936
1.5	0.7726	−0.0105	0.8122	−0.0344	0.7323	+0.0138
1.6	0.7557	+0.0404	0.8304	+0.0302	0.6797	+0.0510

(continued)

UMS	Overall accuracy	Overall accuracy difference	Absent class accuracy	Absent class accuracy difference	Present class accuracy	Present class accuracy difference
2.1.1	0.8265	+0.0748	0.9379	+0.0125	0.5352	−0.0219
2.1.2	0.8228	+0.0779	0.9365	+0.0123	0.5261	−0.0119
2.1.3	0.8170	+0.0805	0.9414	+0.0088	0.4921	−0.0188
2.1.4	0.8315	+0.0777	0.9431	+0.0328	0.5401	−0.0889
2.1.5	0.8296	+0.0813	0.9480	+0.0310	0.5206	−0.0847
2.2	0.8305	+0.0766	0.9462	+0.0285	0.5281	−0.0906
2.3	0.8308	+0.0782	0.9465	+0.0278	0.5288	−0.0793
3.1	0.8309	+0.0762	0.9427	+0.0283	0.5390	−0.0796
4.1.1	0.8022	−0.0070	0.9159	−0.0357	0.5728	+0.0525
4.1.2	0.8768	+0.0191	0.9760	−0.0037	0.3591	+0.1192
4.2.1 (optimal)	0.8065	−0.0026	0.9237	−0.0002	0.5707	−0.0048
4.2.2	0.8800	+0.0130	0.9703	+0.0044	0.4096	+0.0402
4.3.1	0.8060	−0.0052	0.9216	−0.0040	0.5734	−0.0053
4.3.2	0.8793	+0.0139	0.9720	+0.0052	0.3962	+0.0417
4.4.1	0.8066	−0.0053	0.9226	−0.0034	0.5731	−0.0068
4.4.2	0.8799	+0.0136	0.9721	+0.0041	0.3994	+0.0455

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. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (accessed 8 February 2024).

Austen, Martha. 2017. “Put the groceries up”: Comparing black and white regional variation. *American Speech* 92(3). 298–320.

Barreda, Santiago. 2021. Fast Track: Fast (nearly) automatic formant-tracking using Praat. *Linguistics Vanguard* 7(1). 20200051.

Bartlett, Christopher. 2002. *The Southland variety of New Zealand English: Postvocalic /r/ and the BATH vowel*. Otago: University of Otago PhD thesis.

Becker, Kara. 2009. /r/ and the construction of place identity on New York City’s Lower East Side. *Journal of Sociolinguistics* 13(5). 634–658.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major & Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? 🦜 In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’21)*, 610–623. New York: Association for Computing Machinery.

Bender, Emily M. & Alvin Grissom. 2024. Power shift: Toward inclusive natural language processing. In Anne H. Charity Hudley, Christine Mallinson & Mary Bucholtz (eds.), *Inclusion in linguistics* (Oxford Collections in Linguistics), 199–224. Oxford: Oxford University Press.

Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns & Aaron Roth. 2021. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* 50(1). 3–44.

Blodgett, Su Lin & Brendan O’Connor. 2017. Racial disparity in natural language processing: A case study of social media African-American English. Paper presented at the 2017 Workshop on Fairness, Accountability, and Transparency in Machine Learning, Halifax, Nova Scotia, Canada, 14 August. arXiv.

Boersma, Paul & David Weenink. 2022. Praat: Doing phonetics by computer, version 6.3.01 [Computer program]. Available at: http://www.praat.org/.

Brand, James, Jen Hay, Lynn Clark, Kevin Watson & Márton Sóskuthy. 2021. Systematic co-variation of monophthongs across speakers of New Zealand English. *Journal of Phonetics* 88. 101096.

Breiman, Leo. 2001. Random forests. *Machine Learning* 45(1). 5–32.

Charity Hudley, Anne H., Aris Moreno Clemons & Dan Villarreal. 2023. Language across the disciplines. *Annual Review of Linguistics* 9. 253–272.

Cheshire, Jenny. 2004. Sex and gender in variationist research. In J. K. Chambers, Peter Trudgill & Natalie Schilling-Estes (eds.), *The handbook of language variation and change*, 423–443. Oxford: Blackwell.

Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel & Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *KDD ’17: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 797–806. New York: Association for Computing Machinery.

Dormann, Carsten F., Jane Elith, Sven Bacher, Carsten Buchmann, Gudrun Carl, Gabriel Carré, Jaime R. García Marquéz, Bernd Gruber, Bruno Lafourcade, Pedro J. Leitão, Tamara Münkemüller, Colin McClean, Patrick E. Osborne, Björn Reineking, Boris Schröder,

Andrew K. Skidmore, Damaris Zurell & Sven Lautenbach. 2013. Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 36(1). 27–46.
Duncan, Daniel. 2021. Using hidden Markov models to find discrete targets in continuous sociophonetic data. *Linguistics Vanguard* 7(1). 20200057.
Dunn, Jonathan. 2022. *Natural language processing for corpus linguistics*. Cambridge: Cambridge University Press.
Eckert, Penelope & Sally McConnell-Ginet. 1992. Think practically and act locally: Language and gender as community-based practice. *Annual Review of Anthropology* 21. 461–490.
Field, Anjalie, Su Lin Blodgett, Zeerak Waseem & Yulia Tsvetkov. 2021. A survey of race, racism, and anti-racism in NLP. arXiv:2106.11410 [cs]. Available at: https://doi.org/10.48550/arXiv.2106.11410.
Fosler-Lussier, Eric, Laura Dilley, Na'im R. Tyson & Mark A. Pitt. 2007. The buckeye corpus of speech: Updates and enhancements. *Proceedings of Interspeech* 8. 934–937.
Fromont, Robert & Jennifer Hay. 2012. LaBB-CAT: An annotation store. In *Proceedings of Australasian language technology association workshop*, 113–117. Dunedin, New Zealand. Available at: https://aclanthology.org/U12-1015.
Gordon, Elizabeth, Lyle Campbell, Jennifer Hay, Margaret Maclagan, Andrea Sudbury & Peter Trudgill. 2004. *New Zealand English: Its origins and evolution*. Cambridge: Cambridge University Press.
Hall-Lew, Lauren & Sonya Fix. 2012. Perceptual coding reliability of (L)-vocalization in casual speech data. *Lingua* 122(7). 794–809.
Hastie, Trevor, Robert Tibshirani & Jerome Friedman. 2009. *The elements of statistical learning: Data mining, inference, and prediction*. Berlin: Springer.
Heselwood, Barry. 2009. Rhoticity without F3: Lowpass filtering and the perception of rhoticity in "NORTH/FORCE", "START", and "NURSE" words. *Leeds Working Papers in Linguistics and Phonetics* 14. 49–64.
Holliday, Nicole R. 2019. Multiracial identity and racial complexity in sociolinguistic variation. *Language and Linguistics Compass* 13(8). e12345.
Holliday, Nicole & Dan Villarreal. 2020. Intonational variation and incrementality in listener judgments of ethnicity. *Laboratory Phonology* 11(1). 1–21.
Irwin, Ruth Beckey. 1970. Consistency of judgments of articulatory productions. *Journal of Speech and Hearing Research* 13(3). 548–555.
Jackson, Steven J., Sarah Gee & Jay Scherer. 2009. Producing and consuming masculinity: New Zealand's (Speight's) "Southern Man". In L. Wenner & S. Jackson (eds.), *Sport, beer, and gender: Promotional culture and contemporary social life*, 181–201. Zurich: Peter Lang.
Kendall, Tyler, Charlotte Vaughn, Charlie Farrington, Kaylynn Gunter, Jaidan McLean, Chloe Tacata & Shelby Arnson. 2021. Considering performance in the automated and manual coding of sociolinguistic variables: Lessons from variable (ING). *Frontiers in Artificial Intelligence* 4(43). 648543.
Kleinberg, Jon, Sendhil Mullainathan & Manish Raghavan. 2017. Inherent trade-offs in the fair determination of risk scores. In Christos H. Papadimitriou (ed.), *8th Innovations in Theoretical Computer Science Conference*, vol. 43, 1–23. Germany: Dagstuhl.
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* 117(14). 7684–7689.
Kuhn, Max. 2022. Caret: Classification and regression training, version 6.0.93 [R package]. Available at: https://cran.r-project.org/package=caret.
Labov, William. 1990. The intersection of sex and social class in the course of linguistic change. *Language Variation and Change* 2(2). 205–254.
Labov, William. 2001. *Principles of linguistic change*, vol. 2, *Social factors*. Malden, MA: Blackwell.
Labov, William, Sharon Ash & Charles Boberg. 2006. *The atlas of North American English: Phonetics, phonology and sound change*. Berlin: Mouton de Gruyter.
Lawson, Eleanor, James Scobbie & Jane Stuart-Smith. 2014. A socio-articulatory study of Scottish rhoticity. In Robert Lawson (ed.), *Sociolinguistics in Scotland*, 53–78. London: Palgrave Macmillan.
Lawson, Eleanor, Jane Stuart-Smith & James Scobbie. 2018. The role of gesture delay in coda /r/ weakening: An articulatory, auditory and acoustic study. *Journal of the Acoustical Society of America* 143(3). 1646–1657.
Markl, Nina. 2022. Language variation and algorithmic bias: Understanding algorithmic bias in British English automatic speech recognition. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, 521–534. New York: Association for Computing Machinery.
Matsuki, Kazunaga, Victor Kuperman & Julie A. Van Dyke. 2016. The random forests statistical technique: An examination of its value for the study of reading. *Scientific Studies of Reading* 20(1). 20–33.
McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner & Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using Kaldi. In *Proceedings of 18th interspeech*, 498–502.
McLarty, Jason, Taylor Jones & Christopher Hall. 2019. Corpus-based sociophonetic approaches to postvocalic r-lessness in African American language. *American Speech* 94. 91–109.
Mengesha, Zion, Courtney Heldreth, Michal Lahav, Juliana Sublewski & Elyse Tuennerman. 2021. "I don't think these devices are very culturally sensitive" – impact of automated speech recognition errors on African Americans. *Frontiers in Artificial Intelligence* 4. 725911.
Nagy, Naomi & Patricia Irwin. 2010. Boston (r): Neighbo(r)s nea(r) and fa(r). *Language Variation and Change* 22(2). 241–278.
Nguyen, Dong, Dolf Trieschnigg, A. Seza Doğruöz, Rilana Gravel, Mariët Theune, Theo Meder & Franciska De Jong. 2014. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In *COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 1950–1961. Dublin. Available at: https://aclanthology.org/C14-1184.

Pitt, Mark A., Keith Johnson, Elizabeth Hume, Scott Kiesling & William Raymond. 2005. The buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication* 45(1). 89–95.

Podesva, Robert J. 2007. Phonation type as a stylistic variable: The use of falsetto in constructing a persona. *Journal of Sociolinguistics* 11(4). 478–504.

Purse, Ruaridh. 2019. The articulatory reality of coronal stop "deletion". In Sasha Calhoun, Paola Escudero, Marija Tabain & Paul Warren (eds.), *Proceedings of 19th ICPhS*, 1595–1599. Canberra, Australia: Australasian Speech Science and Technology Association. Available at: https://assta.org/proceedings/ICPhS2019/.

R Core Team. 2022. *R: A language and environment for statistical computing, version 4.2.0*. Vienna: R Foundation for Statistical Computing. Available at: https://www.R-project.org/.

Strobl, Carolin & Achim Zeileis. 2008. Danger: High power! Exploring the statistical properties of a test for random forest variable importance. In *COMPSTAT 2008: Proceedings of the 18th International Conference on Computational Statistics*. Porto, Portugal.

Stuart-Smith, Jane. 2007. A sociophonetic investigation of postvocalic /r/ in Glaswegian adolescents. In J. Trouvain & W. J. Barry (eds.), *Proceedings of the 16th International Congress of Phonetic Sciences*, 1449–1452. Saarbrücken: University of Saarbrücken.

Stuart-Smith, Jane, Eleanor Lawson & James Scobbie. 2014. Derhoticisation in Scottish English: A sociophonetic journey. In Chiara Celata & Silvia Calamai (eds.), *Advances in sociophonetics*, 59–96. Amsterdam: John Benjamins.

Tagliamonte, Sali A. & R. Harald Baayen. 2012. Models, forests, and trees of York English: *Was/were* variation as a case study for statistical practice. *Language Variation and Change* 24(2). 135–178.

Villarreal, Dan. 2023. SLAC-fairness: Tools to assess fairness and mitigate unfairness in sociolinguistic auto-coding. Available at: https://djvill.github.io/SLAC-Fairness/.

Villarreal, Dan, Lynn Clark, Jennifer Hay & Kevin Watson. 2019. How to train your classifier. Available at: https://nzilbb.github.io/How-to-Train-Your-Classifier/How_to_Train_Your_Classifier.

Villarreal, Dan, Lynn Clark, Jennifer Hay & Kevin Watson. 2020. From categories to gradience: Auto-coding sociophonetic variation with random forests. *Laboratory Phonology* 11(6). 1–31.

Villarreal, Dan, Lynn Clark, Jennifer Hay & Kevin Watson. 2021. Gender separation and the speech community: Rhoticity in early 20th century Southland New Zealand English. *Language Variation and Change* 33(2). 245–266.

Wassink, Alicia Beckford, Rob Squizzero, Campion Fellin & David Nichols. 2018. Client Libraries Oxford (CLOx): Automated transcription for sociolinguistic interviews. Available at: https://clox.ling.washington.edu/.

Wright, Marvin N., Stefan Wager & Philipp Probst. 2021. Ranger: A fast implementation of random forests, version 0.14.1 [R package]. Available at: https://cran.r-project.org/package=ranger.

Yaeger-Dror, Malcah, Tyler Kendall, Paul Foulkes, Dominic Watt, Jillian Oddie, Daniel Ezra Johnson & Philip Harrison. 2009. Perception of "r": A cross-dialect comparison. Paper presented at the Linguistic Society of America Annual Meeting, San Francisco, 8–11 January.

Zhou, Xinhui, Carol Y. Espy-Wilson, Suzanne Boyce, Mark Tiede, Christy Holland & Ann Choe. 2008. A magnetic resonance imaging-based articulatory and acoustic study of "retroflex" and "bunched" American English /r. *Journal of the Acoustical Society of America* 123(6). 4466–4481.

Zimman, Lal. 2018. Transgender voices: Insights on identity, embodiment, and the gender of the voice. *Language and Linguistics Compass* 12(8). e12284.