Access to this work was provided by the University of Maryland, Baltimore County (UMBC) ScholarWorks@UMBC digital repository on the Maryland Shared Open Access (MD-SOAR) platform.

Please provide feedback

Please support the ScholarWorks@UMBC repository by emailing scholarworks-group@umbc.edu and telling us what having access to this work means to you and why it's important to you. Thank you.

Measuring Equality in Machine Learning Security Defenses

LUKE E. RICHARDS, University of Maryland, Baltimore County; Pacific Northwest National Laboratory, USA

EDWARD RAFF, University of Maryland, Baltimore County; Booz Allen Hamilton, USA

CYNTHIA MATUSZEK, University of Maryland, Baltimore County, USA

The machine learning security community has developed myriad defenses for evasion attacks over the past decade. An understudied question in that community is: for whom do these defenses defend? In this work, we consider some common approaches to defending learned systems and whether those approaches may offer unexpected performance inequities when used by different sub-populations. We outline simple parity metrics and a framework for analysis that can begin to answer this question through empirical results of the fairness implications of machine learning security methods. Many methods have been proposed that can cause direct harm, which we describe as biased vulnerability and biased rejection. Our framework and metric can be applied to robustly trained models, preprocessing-based methods, and rejection methods to capture behavior over security budgets. We identify a realistic dataset with a reasonable computational cost suitable for measuring the equality of defenses. Through a case study in speech command recognition, we show how such defenses do not offer equal protection for social subgroups and how to perform such analyses for robustness training, and we present a comparison of fairness between two rejection-based defenses: randomized smoothing and neural rejection. We offer further analysis of factors that correlate to equitable defenses to stimulate the future investigation of how to assist in building such defenses. To the best of our knowledge, this is the first work that examines the fairness disparity in the accuracy-robustness trade-off in speech data and addresses fairness evaluation for rejection-based defenses.

1 INTRODUCTION

Systems that integrate machine learning introduce a new attack surface for adversaries to take advantage of in terms of security and fairness. When developing defenses for these models, only a few works take any metric beyond the increase in robustness to attack and the accuracy into account. Many defenses are not adequately evaluated, leading to a false sense of security [5]. This occurrence is due to a difference in evaluating a system versus a model. Many adversarial attacks exist in isolation from the entire pipeline, such as preprocessing or post-processing. The community is slowly recognizing how such pipelines or system views can affect end-users and the general security of a system. Separately there is a community studying how such systems interact in social contexts. Very few works, as covered in our related work section, tackle the machine learning system view and the social context in which these systems are developed.

This work aids this understanding of what questions the community should ask when integrating various methods proposed. The security of machine learning is being recognized and reaching the maturity of deployment in private and public sectors, and this is not a far-out issue. While complex threat model levels of access assumptions exist in the literature, the concepts covered in this work examine models attempting to protect from potential adversarial human inputs. This type of attack represents a level of access only at test time, where the ML system creators have to take steps in development to protect against future threats. This threat model and ML security intervention development motivate this work in evaluation standards for measuring such defenses' unequal social failure cases.

Authors' addresses: Luke E. Richards, lerichards@umbc.edu, University of Maryland, Baltimore County; and Pacific Northwest National Laboratory, USA; Edward Raff, University of Maryland, Baltimore County; and Booz Allen Hamilton, USA; Cynthia Matuszek, University of Maryland, Baltimore County, USA.

In more detail, we analyze scenarios in which a defender attempts to protect against an evasion attack, in which the adversary attempts to obfuscate the actual class label during prediction. This case is a realistic threat for in-production models as both user malicious and non-malicious may test the limits of both adversarial robustness and general robustness (e.g., out-of-distribution samples).

Current adversarial defenses treat the dataset used in the evaluation as a monolithic sample that can be summarized by a single statistic rather than examining the complexity of how such interventions can affect end-users. This evaluation method ignores a recurring real-world phenomenon of inequality that such defenses may present. Within this work, we analyze the cases of disparate defense in robustness training and disparate false rejection in popular proposed defenses. We present the case for a more comprehensive evaluation of defenses to account for their downstream implications, especially as such defenses are deployed in critical safety environments.

Within this work, we expand the concept of fairness or robustness to include the adversarial case in which a defender attempts to protect against attacks rather than studying robustness in a clean environment. Many defenses have been introduced over the past few decades (see Related Work). However, none have evaluated the social bias such defenses could present on realistic datasets with actual social subgroup labels.

We motivate this problem by imagining a scenario in a manufacturing or delivery fulfillment environment where a worker attempts to give a speech command to a robotic operation machine. Due to the safety concern of the incorrect machine operation posing risks to humans, the manufacturer added a rejection-based defense method to avoid acting on adversarial commands. The defense offered here is rejection or abstinence, a standard method introduced in defense literature [9, 10, 27]. The speech command recognition model was never tested with a diverse range of end-users. As such, it systematically rejects men in their twenties with Indian accents as it perceives them as potential adversarial examples.

Notably, such a scenario can illustrate the dire need for an evaluation standard that considers how rejection methods perform on various possible end-users. To our knowledge, no existing work on rejection methods has proposed or performed such an evaluation.

Many have faced technology not recognizing them [4], and here machine learning security methods add another layer of complexity that requires analysis that the community has not yet engaged with. There is a lack of benchmarks and taxonomy for how such defenses can present unfairness. We propose a framework for analyzing this disparity within the taxonomy of current defenses. In our case study in speech command recognition, we showcase how such biases can be found in robustness training (adversarial training, pretraining, and data augmentation), preprocessing (denoising), and detection-based defenses (randomized smoothing and neural rejection).

Our contributions are the following:

- Creating a taxonomy for measuring equality in adversarial defenses in robustness training, rejection, and preprocessing defenses.
- Identifying two metrics that capture behaviors over security levels that can be used to measure group defense parity.
- Identifying a task-based speech command recognition dataset with social labels that avoids the bias of common fairness and robustness evaluations in gender classification through facial features.
- Conducting two case studies: 1) on how robustness training methods impact equality of defenses and 2) a comparison of rejection-based methods and their fairness implications.

2 RELATED WORK

2.1 Machine Learning Security

Machine learning security is a growing field [3]. One of the most prototypical attacks is generating digital adversarial examples to evade a classification model. Many attack methods exist that attempt to be less computationally expensive [32] and transfer better to other models [44]. Along with this, there are numerous defenses being proposed. These include adversarial training, including adversarial examples in the training loop [18], certifying models [9], preprocessing [42], and randomization [19]. While it is out of scope computationally to evaluate each of these methods, we offer a framework such that current and future defense methods can be evaluated with respect to fairness.

2.2 Algorithmic Fairness

Creating models and algorithmic systems that use them to make decisions has been shown time and time again to amplify and crystallize social biases. Algorithmic Fairness encompasses a body of work that addresses measuring and mitigating social biases instilled through algorithms [31]. It would not do justice to this rich literature to attempt to summarize all works in this space, so we focus on the intersection between this body of work and security. Our work focuses on using social labels to form subsets or slices of analysis. Works such as [13] have pioneered this form of analysis to measure better how machine learning models fail and metrics for acceptable failure between subgroups.

2.3 Algorithmic Fairness and Security

Adversarial machine learning has been shown to be a beneficial tool in learning less biased representations from data [39, 43]. On the other hand, adversarial methods have been used to degrade the fairness of a model purposefully. Poisoning attacks have been introduced, which further social group disparity in equalized odds between a privileged and unprivileged group [24, 34]. More complex poisoning attacks exist that focus on damaging a sub-population's performance while preserving all others [23]. Frameworks have been introduced to attempt to understand the effect of poisoning attacks where the adversary manipulates subgroup labels on accuracy and fairness [7].

In contrast to our work, they examined how current defenses against poisoning attacks perform. Works have also investigated subverting social attribute re-ranking in image search by modifying the image database with targeted attribute detection [16]. Other work has examined how similar test-time poisoning attacks can be performed on clustering algorithms [8].

However, works have only rarely examined a model's security in relation to the fairness of social factors. Prior works have examined adversarial hardening and fairness, defined as the vulnerability of each class within a classification task [41] in completely class-balanced datasets. Researchers have begun to address critical issues of imbalanced datasets [38] but are still examining and remediating this for only class level. Similarly, evaluations have been proposed for measuring the robustness of data points through distance from decision boundaries and reliance on high-frequency features [29]. Still, these phenomena are studied in the context of class fairness outside of social contexts. Our work begins to explore the effects on real social groups rather than proxy class labels. The metrics optimized for fairness within these works also include accuracy, which does not account for all defense methods that can have systemic rejection (as shown in Section 3.2).

Our work most closely resembles the work of [28], which introduces the concept of biased robustness. That is, a model has different levels of robustness for sub-populations. Within their work, they attempt to use adversarial examples as an upper bound to measure the needed permutation for an evasion attack to be successful. They defined a

lower bound for robustness through a randomized smoothing certification method. While this work is a critical step for equal robustness, it audits models without a threat model present. We expand upon this concept in our work as we acknowledge that a model creator would desire to take mitigation steps to protect against adversarial attacks. We are then interested in the inequality of these interventions with respect to sub-population robustness. Current works have begun examining these trade-offs between fairness and robustness in tabular data in a binary fairness grouping when applying a single defense intervention of adversarial training [36].

We note many of the works [16, 23, 28] in fairness and adversarial machine learning focus on attribute classification through facial recognition. Many of these focus on gender recognition in a binary and from crowd-sourced perceived gender. While the motivation is of deep importance to make machine learning models robust for all, we suggest that gender recognition is in itself a biased task and distracts from a proper evaluation of the usability of machine learning models on a task by end-users [20]. We thus examine speech applications, as the social attributes used to measure fairness are not entangled with the task itself. We expand upon a dataset that can be used for the future in Section 4.1.

2.4 Adversarial Attacks and Defenses in Speech Domain

In our case study, we examine the equality of defense in the speech recognition domain. Works within this space have shown that methods used in the image domain can transfer to the speech domain in untargeted attacks [17] and targeted phrase attacks [6]. Novel defenses within this space have been introduced through methods of data compression [6], audio denoising/purification [35], and randomized smoothing [42]. These studies view the dataset as a homogeneous group and report solely dataset-level performance.

3 FRAMEWORK

We propose a framework to analyze the bias of commonly used defenses in machine learning against adversarial interactions. We consider a model f with parameters θ that takes inputs x and outputs them to some space y. Our threat model is an adversary producing a modified input x' to make the model f exhibit undesired behavior. Within our framework, we analyze the worst-case scenario in which an adversary has both the model class and parameters such that they can perform a white-box attack. This gives an upper bound on the success of an attack and the potential vulnerabilities.

We extend the definition of vulnerability such that we have for some subset of X in the evaluation set labels S that correspond to different social sub-populations of humans. We define biased defenses as there being systemic disparity protections under attack or the rejection of clean examples from a subgroup $s \in S$.

We differ from prior work examining biased robustness in that we assume that there is an adversary attempting to cause general, non-biased attacks on our model, such that the model creator takes steps to either 1) perform robustness training such that the model *should* be robust against an attack or 2) reject the data point, in which a function attempts to categorize whether the data point is an attack. Primarily due to properties of robustness training, a model may have parity of biased robustness but would not capture the cost of overall accuracy for a subgroup s. The introduced metrics in the following sections help capture the trade-off between accuracy, security, and fairness of security interventions.

3.1 Measurements of Biased Vulnerability in Robustness Training and Preprocessing Defense Methods

Robustness training can take many forms. This may look like data augmentation, adversarial training [26], large-scale pretraining with datasets attempting to capture diversity [14], and a mixture of all methods. Adversarial training offers the most well-studied empirical defense to adversarial attacks [26]. This method involves including attacked

examples during training with the goal of classifying attacked examples correctly, $f(x + \delta) = y$ despite the adversaries' optimization of $f(x + \delta) \neq y$. The method for finding such a δ has varied, with many proposals for faster and more efficient methods [12, 32].

Due to the smaller scale of our experiment, we can choose to use a strong attack during training of projected gradient descent (PGD) [26]. It has been shown that most methods attempt to approximate this optimization [15]. Learning stability is critical in the accuracy-robustness trade-off, and thus choosing larger attack budgets for ϵ can have weaker general *and* attacked performance. Thus we must choose a realistic epsilon value. However, to our knowledge, this is the first work that examines how the fairness of this accuracy-robustness trade-off disparately impacts real social groups in a non-tabular dataset.

We examine the differences between the regularly trained model and models with various robustness defense interventions across the various sub-populations. We measure accuracy across sub-populations, leading to measuring the Accuracy-Parity (AP). We examine the disparity that such defense interventions introduce through ablation studies. We note this metric should reflect the performance metric for the task, such as word error rate (WER) for speech recognition or equal error rate (EER) for speaker verification.

We measure the area under the curve (AUC) for samples of varying levels of attack budgets $[\epsilon_{min}, \epsilon_{max}]$ and the resulting accuracy for a slice of data (X_s, Y_s) where s is the subgroup Eq. 1. This summary statistic of accuracy area under the curve accounts (AUC_{acc}) for the early loss of performance for subgroups when introducing a defense. Here a lower AUC_{acc} indicates less protection under attack.

$$AUC_{acc}(X, Y, s) = \int_{\epsilon_{min}}^{\epsilon_{max}} \frac{[f(X_s + \delta_\epsilon) == Y_s]}{|X_s|} \, dx. \tag{1}$$

We are interested in understanding how these interventions cause larger gaps in the defense from interventions. To address this, we introduce the Defense Parity (DP) metric, which intuitively attempts to capture the difference in defense performance across different subgroups of potential users. We define Defense Parity as the largest difference between subgroups as we apply defense interventions Eq. 2. We can extend this analysis for preprocessing defenses that attempt to cleanse adversarial examples during test time. Such methods attempt to solve $f(g(x + \delta)) = y$ by introducing a method g, which can take many forms.

$$DP = \max_{s_i, s_j \in S} [AUC_{acc}(X, Y, s_i) - AUC_{acc}(X, Y, s_j)]. \tag{2}$$

3.2 Measurements of Biased Rejection Defense Methods

When accounting for the rejection of a sample through a mechanism for classifying adversarial examples, the bias can be measured using the false positivity rate (FPR). We measure the FPR for each group $s \in S$ as a clear metric for whether a single group is being wrongly flagged more often as an adversarial attack. The simplicity of this approach extends to not needing an adversary present in evaluation, thereby limiting the technical implementation bias of attack strength or method. This is particularly appealing as many flaws can be found in how current defenses are evaluated [5, 15].

We define f^* as a classifier that produces $\hat{y}$ the class prediction or an abstain signal (-1). We measure the AUC of FPRs for each group $s \in S$ with corresponding data X_s over threshold values $[\alpha_{min}, \alpha_{max}]$ (Eq. 3). We then examine the FPR parity between subgroups by measuring the largest difference between subgroups (Eq. 4). These metrics capture an ideal case with a stable increase as the security threshold α increases for all groups $s \in S$. The goal here is to decrease the AUC_{FPR} in general while also increasing the parity.

$$AUC_{FPR}(X, s) = \int_{\alpha_{min}}^{\alpha_{max}} \frac{f^*(X_s) == -1}{|X_s|} \mathrm{d}x. \tag{3}$$

$$FPRP = \max_{s_i, s_j \in S} [AUC_{FPR}(X, s_i) - AUC_{FPR}(X, s_j)]. \tag{4}$$

Multiple rejection-based methods have been proposed within the adversarial machine learning literature. However, no work so far has examined this rejection mechanism in the context of social subgroup bias. For our case studies, we compare two approaches that use different methods for determining rejection. These metrics allow us to understand first the general overall performance for each subgroup and then the overall equality of defense for groups. We note that high FPR parity with high false rejections AUC_{FPR} may occur and may be desired if no alternatives exist. Ideally, the algorithm should provide high FPR parity with low false rejection scores per group AUC_{FPR}.

4 EXPERIMENTS

In this section, we outline our experiments in the domains of keyword recognition speech models. We cover the methods by which we harden each model through adversarial training and the methods we use for a rejection method.

4.1 Dataset

For an exhaustive evaluation of this framework, we use a shorter subset of Common Voice [1], which has single-word phrases in English (retrieved June 2022, version 9). We choose this subset for being 1) lightweight and offering more minor compute costs for analyzing the equality of defenses and 2) for being a classification dataset rather than a sentence-level automatic speech recognition system, which has a higher cost in computation and expertise to train. Classification, in particular, has been the only task studied in rejection methods. Adapting such methods would be an effort in itself. As well, the standard classification tasks on non-tabular data are typically facial trait recognition (gender, age, etc.) which has a high existing social bias (Section 2.3).

The size of the dataset also allows researchers to run more extensive exhaustive empirical experiments regarding computing-intensive methods such as adversarial training. While datasets such as [21] offer the criterion of socially labeled data, it is computationally expensive to process out just the audio and, again, is a more complex task of speech-to-text.

We focus on the English subset, which has 15,115 training examples, 7,634 validation examples, and 7,640 test examples. We remove all examples of the command Firefox, leaving 12 classes total ("hey", "yes/no", and the numbers 0 through 9). Common Voice allows contributors to self-identify their age, gender (labeled as the typical term for sex within the dataset), and accent. The gender ratio within the train set is 25% female, 71% male, and 2% other. There are 14 self-identified English accents in the train/test set. The majority of the dataset is represented by those identifying as United States English accents (47%) with English (England) (17%) and Indian English (12%) accents following. Age distributions are heavily weighted to people in their twenties (42%) and few examples for those in their sixties (4%), seventies (1%), and eighties (0.1%), with all other age groups having between 9% and 16%.

This makes the data biased toward younger US English speakers. While this lack of representation in the dataset is not ideal, this is a typical case for datasets that are currently driving models within speech today. When examined at the intersection of gender, accent, and age, the dataset contains 87 unique groups in the training set and 49 in the test set. There are 14 unique groups not present in the training set but within the test set. This introduces an interesting problem of generalization to variations based on just the intersections that are in the training set. The average length

of a file is 0.0208 ± 0.0001 milliseconds. The collection was standardized on Common Voice's platform, with each file having a sampling rate of 48 kHz.

4.2 Biased Vulnerability in Adversarial Robustness Defenses Study

We focus our study on a one-dimensional convolutional network modeled after the M5 architecture [11]. We also evaluate a version of M5 with a handful of tricks shown to increase adversarial robustness by removing batch normalization [37] and by using a different activation function SiLU [22]. We call this version M5-Tricks. We also explore wav2vec 2.0 [2], a large self-supervised pretrained speech recognition model.

We down-sample from the original sample rate of 48 kHz to 16 kHz. We run ablations on adding data augmentation as prior work [33] has found it to be essential for accuracy-robustness trade-off; for augmentations, we add Gaussian noise with a mean of 0 and varying standard deviation of $\{0, 0.1, 0.3, 0.5\}$. We do not perform augmentations during validation to ensure we choose the best clean performance model. We perform ablation studies by training a model with and without adversarial training and with and without noise augmentation. For every study, we analyze the subgroups of accent, age, and gender.

For adversarial training of the model, we incorporate adversarial examples crafted by a PGD attack with $\epsilon = \{0.01, 0.1\}$, 10 steps, and an alpha of 5 times lower than epsilon ($\alpha = \epsilon/5$). We add adversarial examples in the same batch as our clean examples and optimize for the joint loss with $\lambda = .5$, weighting both the adversarial and clean performance equally. We perform the same evaluation on the validation set for model selection to optimize for the model's clean and attacked accuracy.

For our studies, we analyze three classes of models, the previously described CNN, a version of the M5 model with a handful of tricks for adversarial training, wav2vec 2.0, and models with and without noise augmentations. We use the HuggingFace transformers library [40] for loading the wav2vec 2.0 model.

We fine-tune the large internet-scale model on the Common Voice Clips dataset by freezing the encoder model and training a multi-layer perception on top. We fine-tune multiple versions of the wav2vec 2.0 model with varying levels of random noise augmentation and without. Every model is trained for 1000 epochs, with an Adam optimizer [25] with an initial learning rate of 1e-3 that decreases at a rate of 0.90 every 5 epochs using cross-entropy loss. For the wav2vec 2.0 models, we use a learning rate scheduler that decreases at 0.50 every 25 epochs. We select the model with the lowest loss on the validation set depending on the training regime. For our adversarial robustness evaluation, we run attacks on all models with a range of attack budget strengths $\epsilon = \{0.0001, 0.001, 0.01, 0.1, 0.2, 0.3\}$ and with steps $\{1, 10, 50\}$. Our learning rate $\alpha = \epsilon/5$.

In summary, we examine how the equality of defenses is impacted by noise augmentation with a standard deviation of 0.1, 0.3, and 0.5 (NA1, NA3, NA5), adversarial training tricks (T), adversarial training with budgets of 0.01 or 0.1 (AT.01, AT.1), and large-scale pre-training (wav2vec2). Our resulting training runs leave us with 16 models: M5, M5-NA1, M5-NA3, M5-NA5, M5-T, M5-T-AT.01, M5-T-AT.01NA1, M5-T-AT.01-NA3, M5-T-AT.1, M5-T-AT.1-NA1, M5-T-AT.1-NA1, M5-T-AT.1-NA3, wav2vec2, wav2vec2-NA1, wav2vec2-NA3, and wav2vec2-NA5.

4.3 Biased Rejection in Defenses Study

We compare two proposed methods in the literature for their biased rejection based on the measurement of false positivity rate. We use the M5 model architecture and training procedure outlined above without adversarial training or noise augmentation unless otherwise specified.

For our first rejection method, we analyze the work of [27] by doing neural rejection (NR). Within this framework, it is assumed we have a deep neural model f which has layers $l_0, ..., l_N$, each taking in the previous representation h. It is also assumed that the final layer, l_N, is a fully connected linear layer, which takes a learned representation h_{N-1}. They propose learning a Support-Vector Machine (SVM) with an RBF kernel using h_{N-1}. Due to the SVM being a Compact Abating Probability (CAP) model [27], we can use the probabilities as a measurement from the training distribution. This allows rejecting based on some threshold t, assuming it is an adversarial example. We measure the equalized odds by measuring the percentage per each sub-population in $s \in S$, which are rejected at varying thresholds σ.

The second method we analyze within the framework is randomized smoothing (RS) [9]. This method has been widely adopted as it offers certifiable robustness with a simple implementation. Randomized smoothing can be defined as adding noise ϵ to the classification $f(x + \epsilon) = y$ where ϵ is drawn from $N(0, \sigma^2 I)$. The intuition is that adversarial examples at test time would be 'drowned out' with this noise addition, and the model is trained with such noise. The original work offers a simple algorithm that includes a reject from classifying task based on a Monte Carlo sampling. Here n noise profiles are sampled from $N(0, \sigma^2 I)$ and the prediction classes for each data point with noise profile addition are accumulated as counts. The top two class counts are then used to parameterize a binomial test with a threshold α.

While prior work [28] has used randomized robustness certification as a lower-bounded for robustness, we instead test the rejection or choice to abstain classification. We measure the number of examples from a subgroup that abstained from classification for a smoothed model. Again, we measure the FPR of rejection in case no adversarial examples are present. We use both studies' areas under the curve as a summary statistic. Again our assumption is that a lower AUC_{FPR} indicates a lower FPR rate for a subgroup. Our ideal case would be that given varying levels of security (thresholds), we would see equal rejection levels per subgroup (FPR parity).

For our NR rejection methods, we use the method of [27], with an SVM with an RBF kernel implemented in sklearn [30]. We use the final layer representation from the models to train these rejection models that can also perform classification. For our studies on rejection with RS, we train two M5 networks with $\sigma \in 0.1, 0.3$. We choose models based on a held-out validation set also augmented with the same noise level. Similar to [42], we found that 0.1 began to see lower accuracy and thus only report on $\sigma = 0.1$. We then use varying numbers of samples $N = \{10, 100, 1000, 10000\}$ for the binomial test. This adds an additional parameter that neural rejection does not offer, which we use to explore the implications of fairness of defense. For both methods, we evaluate the α values between $[0.001, 1]$ with a step size of $1e - 3$. This gives us a high-fidelity view of the behavior of false positive rejection across security levels through the use of thresholds. We then compare the two methods on fair rejection parity and report our findings.

5 RESULTS

5.1 Robust Training: Interventions

We report the general or average accuracy over attack budgets for each model (see appendix Figure 6) to visually compare models' performance under attack. We then use the AUC_{acc} summary statistic over attack strength to measure the general defense of each method and the clean accuracy. Here we see that the pretrained model (wav2vec 2.0) had the highest AUC_{acc}. These values are much higher than the second largest of non-pre-trained models M5-AT with the noise of 0.1.

We use Pearson correlation to measure how security interventions have implications for the fairness of subgroups. We encode all interventions as a binary value (intervention applied (1) or not (0)) and report how these methods

Table 1. Randomized Smoothing (RS) AUC_{FPR} Results

Number of Samples	10	100	1000	10000
AUC_{FPR}	0.116	0.029	0.007	0.001

correlate with higher defense measured by AUC_{acc} per each subgroup. We report these results through a heat map in Figure 1. Overall we find general trends in most defense inventions. We find that pretraining is correlated with higher levels of defense for all groups except those with Australian English accents and those in their seventies.

Interestingly, these two groups are some of the few that see a positive correlation with adversarial training (AT). Those in their seventies are the only group with a positive value for this correlation of 0.47. While we see with adversarial tricks (the changing of activation function and removal of batch-norm), we see that the value is consistent with adversarial training but a slight increase in the general direction of the correlation. We see this again with the increase of adversarial budget increasing the robustness for those in their seventies, Australian, and Philippine English accents (Figure 2).

On the broader question of defense parity of methods, we break down the results by reporting the correlation coefficients between defense parity and the intervention methods (Figure 3) and their level of the application when appropriate (Figure 4). We see that adversarial training is correlated with a decrease in defense parity between gender and accent but near zero for age. When examining the level rather than the binary of whether adversarial training was applied, there are stronger correlations for defense parity decreasing when the adversarial training budget increases. In contrast, here we see that all groups have a stronger correlation. Gender has the highest correlation coefficient among the groups. For the age subgroups, in particular, we see a much higher correlation than in the binary case shedding light on the dynamics of adversarial training's budget needing to be optimized for security and fairness in applications.

For noise augmentation, we had weaker correlation coefficients among the gender and age groups where increasing noise level increased defense parity. This is in contrast with the accent, where we saw a weak correlation where increasing the noise parity decreased defense parity. When breaking the results down to the noise augmentation level, we see a weak correlation between gender and accent for increasing the noise augmentation level and defense parity. For accent, however, there is a weak correlation between decreasing defense parity and increasing the noise level.

Our strongest correlation comes from pretraining (using the wav2vec 2.0 model). We see that accent has the highest correlation with this intervention increasing the defense parity. Gender has a weaker correlation for this increase in parity, while age has a decrease in parity. Large-scale training presents a pivotal way to increase defense parity in most cases, but the underlying data and social group representation of this process may play a role in who sees the benefits when examining the defense parity.

5.2 Rejection: Randomized Smoothing vs. Neural Rejection

Our results with the NR method result in a learned model with an overall accuracy of 0.646 and a AUC_{FPR} of 0.337. For the RS model with $\sigma = 0.1$, we have an accuracy of 0.730 ± 0.003 (averaged over 10 noise profiles runs) and a AUC_{FPR} dependent on the number of samples which we report in Table 1. As hypothesized, the RS sampling gives control over AUC_{acc} that NR does not provide, allowing for much lower AUC_{FPR}. We contrast the parity and their correlating factors for gender, age, and accent.

For the NR method, we have AUC_{FPR} parity values of 0.014 for gender, 0.101 for age, and a larger value of 0.273 for accent. For the RS, we see values that change oversampling values ranging from 0.069 to 0.002 for gender, 0.119 to

Correlation Coefficients for Robustness Training

	AT	AT Tricks	Pretraining	Noise Aug.
female	-0.55	-0.74	0.6	-0.18
male	-0.61	-0.86	0.57	-0.17
other	-0.083	-0.43	0.43	-0.51
african	-0.6	-0.87	0.67	-4e-17
australia	0.02	0.077	-0.22	-0.32
canada	-0.035	-0.13	0.39	-0.16
england	-0.46	-0.71	0.33	-0.27
hongkong	-0.69	-0.53	0.59	0.039
indian	-0.52	-0.75	0.33	-0.17
ireland	-0.45	-0.58	0.67	-0.23
philippines	0.041	-0.16	0.28	0.37
singapore	-0.16	-0.63	0.36	0.25
us	-0.65	-0.88	0.64	-0.22
teens	-0.55	-0.61	0.28	-0.49
twenties	-0.62	-0.83	0.61	-0.18
thirties	-0.57	-0.84	0.66	-0.04
fourties	-0.58	-0.81	0.44	-0.098
fifties	-0.56	-0.86	0.54	-0.21
sixties	-0.43	-0.82	0.68	-0.092
seventies	0.47	0.54	-0.56	-0.33

Fig. 1. Pearson correlation coefficient for the binary values of interventions taken during robustness training.

0.004 for age, and 0.178 to 0.003 for accent. Here we can see the benefit of choosing RS when considering the fairness implications of security methods: we can increase the AUC_{FPR} parity by increasing the number of samples. In all cases, when sampling 1000 samples, we achieved higher parity compared to NR for all subgroup types (Figure 5). While this adds $N_{samples}x|X|$ number of model runs, this can be run in parallel and aggregated later. Future work will be needed to analyze the computational cost and latency this may introduce at the cost of fairness in security.

While sampling seems to increase the parity, we seek to understand better the underlying distributions for the subgroups in the NR method and factors that may contribute to the higher parity. Remember that NR relies on the fact
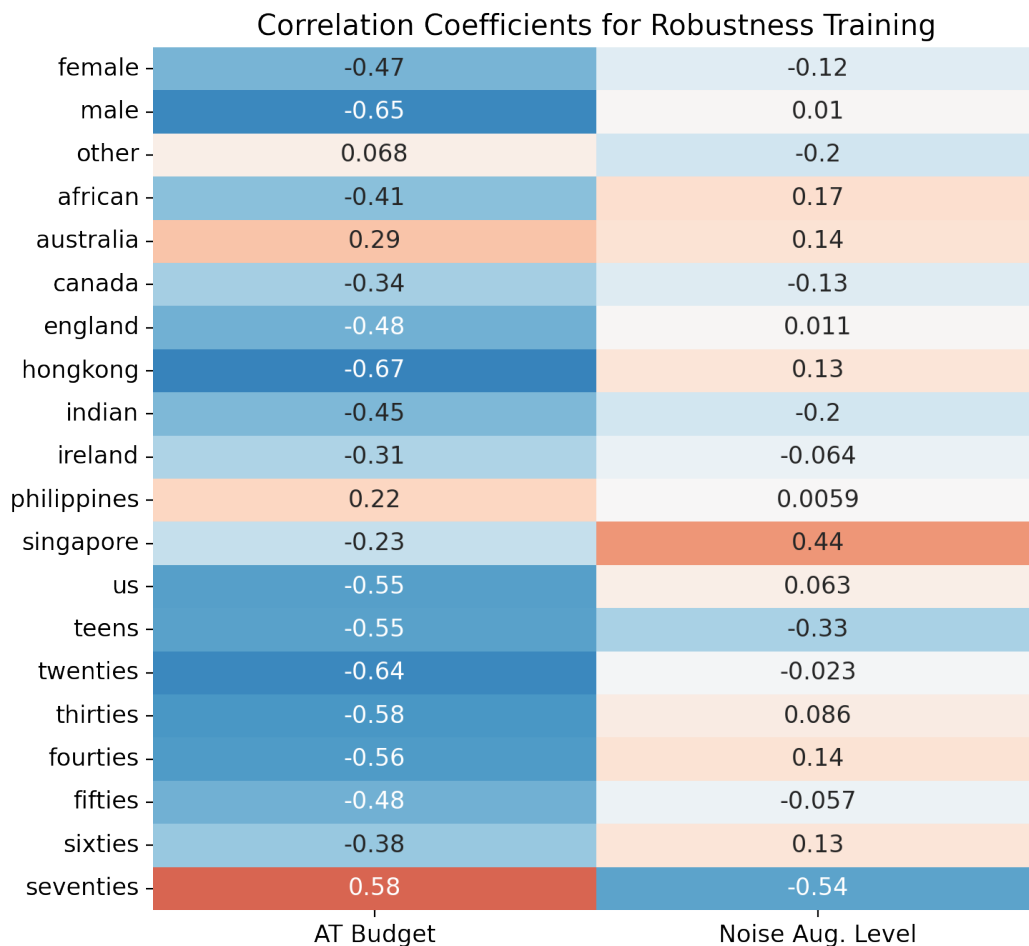
Correlation Coefficients for Robustness Training

	AT Budget	Noise Aug. Level
female	-0.47	-0.12
male	-0.65	0.01
other	0.068	-0.2
african	-0.41	0.17
australia	0.29	0.14
canada	-0.34	-0.13
england	-0.48	0.011
hongkong	-0.67	0.13
indian	-0.45	-0.2
ireland	-0.31	-0.064
philippines	0.22	0.0059
singapore	-0.23	0.44
us	-0.55	0.063
teens	-0.55	-0.33
twenties	-0.64	-0.023
thirties	-0.58	0.086
fourties	-0.56	0.14
fifties	-0.48	-0.057
sixties	-0.38	0.13
seventies	0.58	-0.54

Fig. 2. Pearson correlation coefficient for the continuous values of interventions taken during robustness training.

that the SVM trained on top of the neural network will be a CAP model. Due to this, we investigate if having a higher subgroup representation in training set results in higher AUC_{FPR}. Here we seek to measure the correlation between the size of subgroup training data $|X_{train,s}|$ where s is the subgroup.

For all data combined, we find a correlation coefficient with AUC_{FPR} of -0.032 for the NR method. When calculating this for each subgroup separately, we find gender has a correlation coefficient of 0.691, age has -0.318, and accent has -0.044 for the NR method. Examining the gender groups, there is high parity in AUC_{FPR} where a strong correlation is a result of only having 3 data points (0.309 (M), 0.311 (F), 0.296 (O)). While the intuition for RS is not as straightforward

Correlation between Intervention Methods and Defense Parity
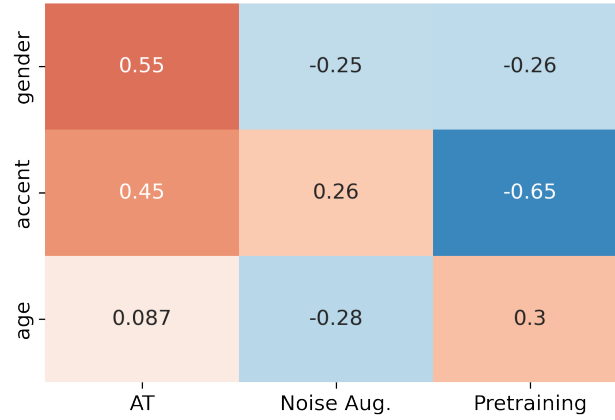
Fig. 3. Pearson correlation coefficient between interventions and their resulting defense parity. A red darker color indicates a positive correlation with lower defense parity and blue with higher defense parity.

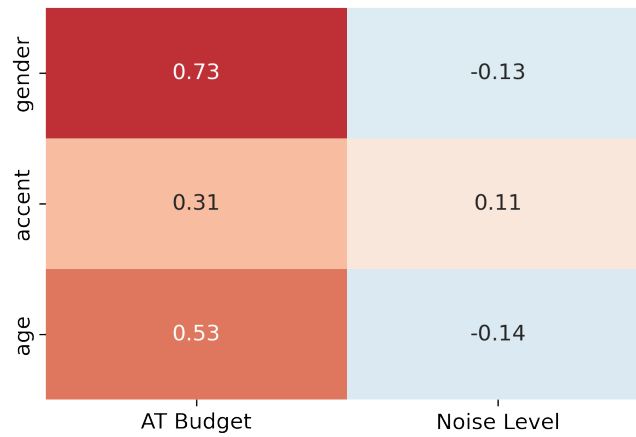Correlation between Intervention Levels and Defense Parity

Fig. 4. Pearson correlation coefficient between noise levels and adversarial training budget interventions and their resulting defense parity. A red darker color indicates a positive correlation with lower defense parity and blue with higher defense parity.

as with NR, we find that RS has more consistent correlations to subgroup training data size. For RS, the correlation with training data size and AUC_{FPR} we see a correlation of -0.499 ± 0.08 (averaged over all sampling). For each subgroup, we see -0.811 ± 0.16 for gender, -0.430 ± 0.144 for age, and -0.652 ± 0.216 for accent. Here we see much stronger trends for having smaller AUC_{FPR} values with increasing training data.

5.3 Discussion

Our framework and case study allows us to return to the original question: Whom do machine learning security methods work? Introducing a rejection method without evaluating in this context would have a system that refuses to recognize
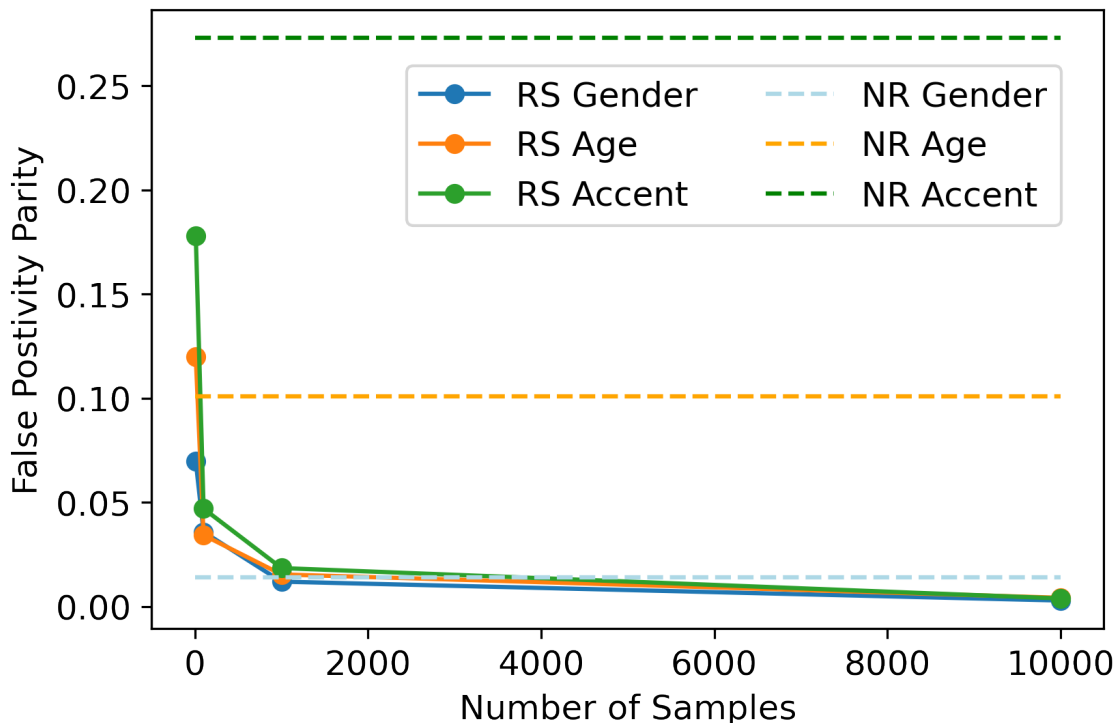
Fig. 5. Parity of the area under the curve for the false positive rate AUC_{FPR} for each method. Randomized smoothing (RS) plotted as solid lines over samples and neural rejection (NR)

speech from older adults and women with Indian English accents (among other groups). Such rejection results in these workers not being able to lose access to a required element of their job. We empirically compared two rejection methods, neural-based rejection, and randomized smoothing rejection, by measuring false positives. We found that the sampling hyperparameter in randomized smoothing allowed for more equality and lowered false positives across all subgroups. This result in itself represents a pivotal contribution to accompany the framework itself.

We analyzed through ablation how various proposed adversarial defense methods protect users under attack from being misunderstood. We found that large-scale pretraining was a consistent increase in robustness but nuance for some participants, with some in age and accent categories seeing a negative correlation. In contrast, adversarial training assisted these subgroups. Both noise and adversarial training demonstrated further nuance as the budget level affected different groups differently. While this work was a first step in addressing the equity of adversarial robustness in the speech domain, there are many open questions future work will use this framework to address.

6 CONCLUSION AND LIMITATIONS

In this work, we introduce an extension of a threat model where security interventions have fairness implications in machine learning systems. We outline the first two parity metrics to help capture these implications. AUC_{acc} (accuracy over attack budgets) measures the parity between subgroups when introducing a robustness training defense and

preprocessing-based defense. AUC_{FPR} (false rejection over levels of security) measures biases in rejection methods that attempt not to classify adversarial attacks. Within our case studies of robustness training, we find this lessens parity for users in the gender and accent subgroups and that large-scale pretraining increases parity but weakly decreases in the case of age. We further this analysis by examining cases where the noise level increases while the level of adversarial budget decreases AUC_{acc} parity.

For our rejection method, we compared randomized smoothing and neural rejection. We achieved higher FPR parity across all groups by increasing the number of samples with randomized smoothing. We also observe much lower false positivity rates in randomized smoothing than in neural rejection. We recommend utilizing methods such as randomized smoothing that allow for multiple axes of configuration (security threshold and the number of samples) to allow for more equal defenses.

We acknowledge that this work relies on the availability of labels for subgroups that can be difficult to obtain in all domains. We also recognize the limitations of such a case study in speech command recognition to not apply directly to other domains. Future work must find benchmarks to address this problem. We hope that by introducing these metrics in this domain and conducting case studies, we can provide awareness of the problem and spur work to address the development and research of equal defenses in machine learning security.

REFERENCES

[1] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670* (2019).

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems* 33 (2020), 12449–12460.

[3] Battista Biggio and Fabio Roli. 2018. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition* 84 (2018), 317–331.

[4] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.

[5] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. 2019. On Evaluating Adversarial Robustness. *arXiv preprint arXiv:1902.06705* (2019).

[6] Nicholas Carlini and David Wagner. 2018. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE security and privacy workshops (SPW)*. IEEE, 1–7.

[7] L Elisa Celis, Anay Mehrotra, and Nisheeth Vishnoi. 2021. Fair classification with adversarial perturbations. *Advances in Neural Information Processing Systems* 34 (2021), 8158–8171.

[8] Anshuman Chhabra, Adish Singla, and Prasant Mohapatra. 2021. Fairness Degrading Adversarial Attacks Against Clustering Algorithms. *arXiv preprint arXiv:2110.12020* (2021).

[9] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. 2019. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*. PMLR, 1310–1320.

[10] Francesco Crecchi, Marco Melis, Angelo Sotgiu, Davide Bacciu, and Battista Biggio. 2022. Fader: Fast adversarial example rejection. *Neurocomputing* 470 (2022), 257–268.

[11] Wei Dai, Chia Dai, Shuhui Qu, Juncheng Li, and Samarjit Das. 2017. Very deep convolutional neural networks for raw waveforms. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 421–425.

[12] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Xiaolin Hu, and Jun Zhu. 2017. Discovering adversarial examples with momentum. *arXiv preprint arXiv:1710.06081* (2017).

[13] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[14] Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yuhao Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. 2022. Data Determines Distributional Robustness in Contrastive Language Image Pre-training (CLIP). *arXiv preprint arXiv:2205.01397* (2022).

[15] Yue Gao, Ilia Shumailov, Kassem Fawaz, and Nicolas Papernot. 2022. On the Limitations of Stochastic Pre-processing Defenses. *arXiv preprint arXiv:2206.09491* (2022).

[16] Avijit Ghosh, Matthew Jagielski, and Christo Wilson. 2022. Subverting Fair Image Search with Generative Adversarial Perturbations. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 637–650. https://doi.org/10.1145/3531146.3533128

[17] Yuan Gong and Christian Poellabauer. 2017. Crafting adversarial examples for speech paralinguistics applications. *arXiv preprint arXiv:1711.03280* (2017).

[18] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations*. http://arxiv.org/abs/1412.6572

[19] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. 2018. Countering Adversarial Images using Input Transformations. In *International Conference on Learning Representations*.

[20] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M Branham. 2018. Gender recognition or gender reductionism? The social implications of embedded gender recognition systems. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–13.

[21] Caner Hazirbas, Joanna Bitton, Brian Dolhansky, Jacqueline Pan, Albert Gordo, and Cristian Canton Ferrer. 2021. Casual conversations: A dataset for measuring fairness in ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2289–2293.

[22] Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415* (2016).

[23] Matthew Jagielski, Giorgio Severi, Niklas Pousette Harger, and Alina Oprea. 2021. Subpopulation data poisoning attacks. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. 3104–3122.

[24] Changhun Jo, Jy-yong Sohn, and Kangwook Lee. 2022. Breaking Fair Binary Classification with Optimal Flipping Attacks. *arXiv preprint arXiv:2204.05472* (2022).

[25] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[26] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* (2017).

[27] Marco Melis, Ambra Demontis, Battista Biggio, Gavin Brown, Giorgio Fumera, and Fabio Roli. 2017. Is deep learning safe for robot vision? Adversarial examples against the icub humanoid. In *Proceedings of the IEEE international conference on computer vision workshops*. 751–759.

[28] Vedant Nanda, Samuel Dooley, Sahil Singla, Soheil Feizi, and John P Dickerson. 2021. Fairness through robustness: Investigating robustness disparity in deep learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 466–477.

[29] Gaurav Kumar Nayak, Ruchit Rawal, Rohit Lal, Himanshu Patil, and Anirban Chakraborty. 2022. Holistic Approach To Measure Sample-Level Adversarial Vulnerability and Its Utility in Building Trustworthy Systems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 4332–4341.

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[31] Dana Pessach and Erez Shmueli. 2020. Algorithmic fairness. *arXiv preprint arXiv:2001.09784* (2020).

[32] Maura Pintor, Fabio Roli, Wieland Brendel, and Battista Biggio. 2021. Fast minimum-norm adversarial attacks through adaptive norm constraints. *Advances in Neural Information Processing Systems* 34 (2021), 20052–20062.

[33] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy A Mann. 2021. Data augmentation can improve robustness. *Advances in Neural Information Processing Systems* 34 (2021), 29935–29948.

[34] David Solans, Battista Biggio, and Carlos Castillo. 2020. Poisoning attacks on algorithmic fairness. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 162–177.

[35] Anirudh Sreeram, Nicholas Mehlman, Raghuveer Peri, Dillon Knox, and Shrikanth Narayanan. 2021. Perceptual-based deep-learning denoiser as a defense against adversarial attacks on ASR systems. *arXiv preprint arXiv:2107.05222* (2021).

[36] Haipei Sun, Kun Wu, Ting Wang, and Wendy Hui Wang. 2022. Towards Fair and Robust Classification. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*. IEEE, 356–376.

[37] Haotao Wang, Aston Zhang, Shuai Zheng, Xingjian Shi, Mu Li, and Zhangyang Wang. 2022. Removing Batch Normalization Boosts Adversarial Training. In *International Conference on Machine Learning*. PMLR, 23433–23445.

[38] Wentao Wang, Han Xu, Xiaorui Liu, Yaxin Li, Bhavani Thuraisingham, and Jiliang Tang. 2021. Imbalanced adversarial training with reweighting. *arXiv preprint arXiv:2107.13639* (2021).

[39] Zhibo Wang, Xiaowei Dong, Henry Xue, Zhifei Zhang, Weifeng Chiu, Tao Wei, and Kui Ren. 2022. Fairness-aware Adversarial Perturbation Towards Bias Mitigation for Deployed Deep Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10379–10388.

[40] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).

[41] Han Xu, Xiaorui Liu, Yaxin Li, Anil Jain, and Jiliang Tang. 2021. To be robust or to be fair: Towards fairness in adversarial training. In *International Conference on Machine Learning*. PMLR, 11492–11501.

[42] Piotr Żelasko, Sonal Joshi, Yiwen Shao, Jesus Villalba, Jan Trmal, Najim Dehak, and Sanjeev Khudanpur. 2021. Adversarial attacks and defenses for speech recognition systems. *arXiv preprint arXiv:2103.17122* (2021).

[43] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 335–340.

[44] Yuekai Zhang, Ziyan Jiang, Jesús Villalba, and Najim Dehak. 2020. Black-Box Attacks on Spoofing Countermeasures Using Transferability of Adversarial Examples.
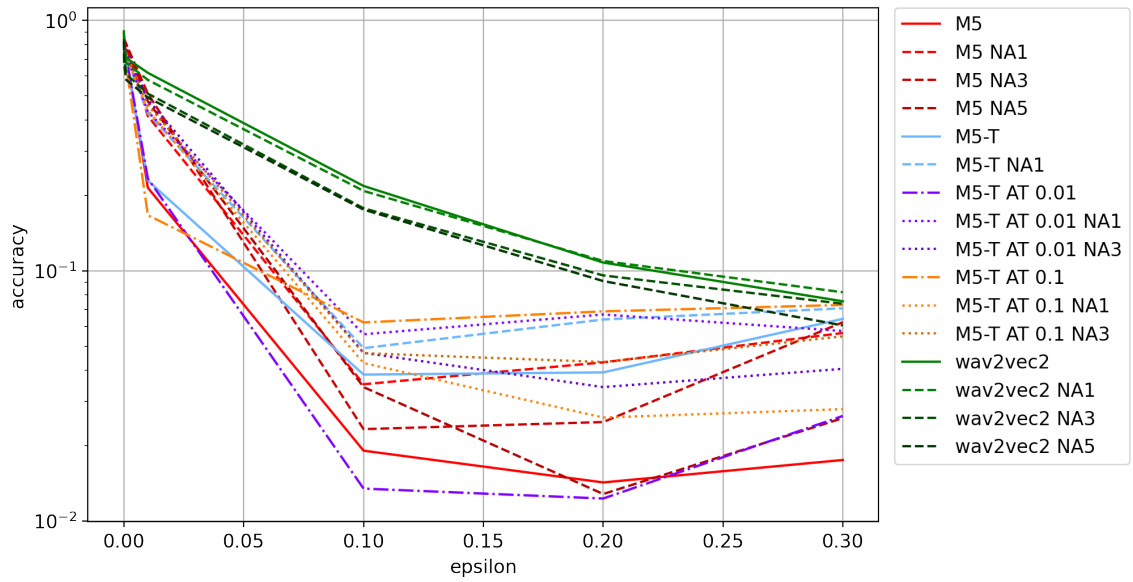
A EXTENDED RESULTS

Fig. 6. The general performance averaged over all subgroups of each model. This serves as the basis for our fairness AUC measured in later parts. We can see that the pretrained large-scale models have overall higher performance overall attack strengths. We see that noise augmentation does not assist the large models until a larger attack strength. In terms of smaller models, we see that the M5 model with the adversarial training tricks and a budget of .1 (M5-T AT .1) performs almost as well when under the highest attack strength.
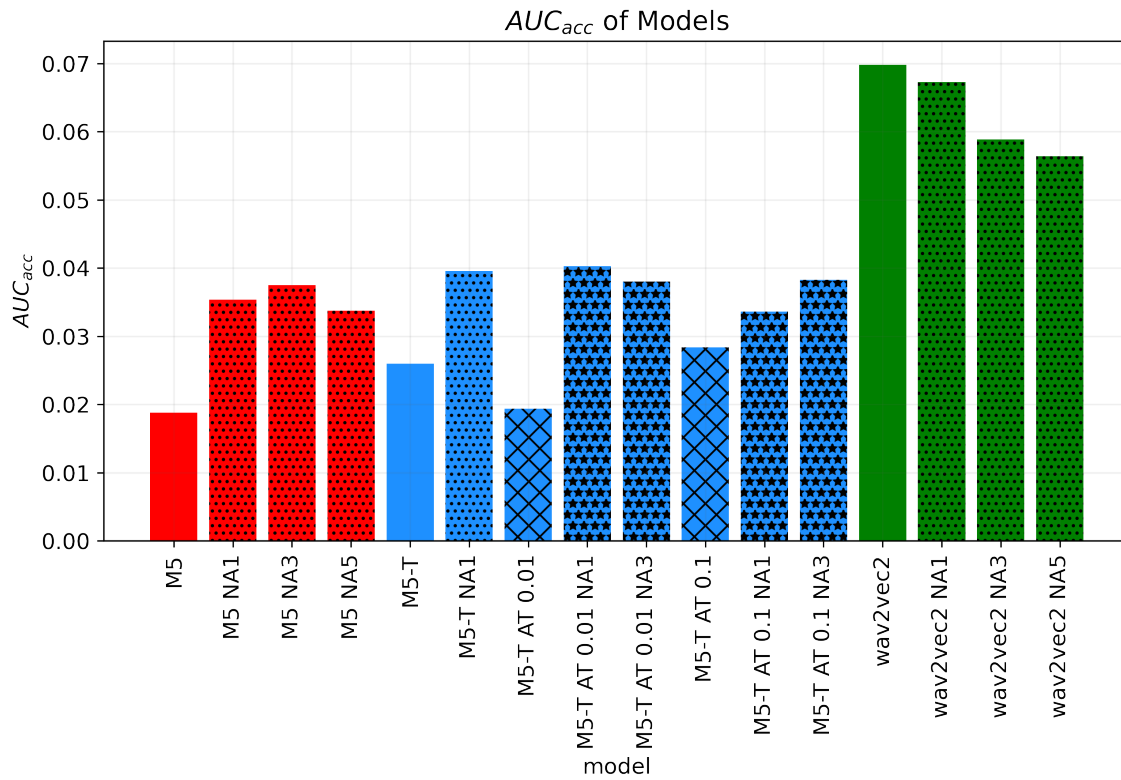
Fig. 7. Overall AUC_{acc} for each model. Dotted indicate noise augmentation, crossed lines indicate adversarial training, and stars indicate both.
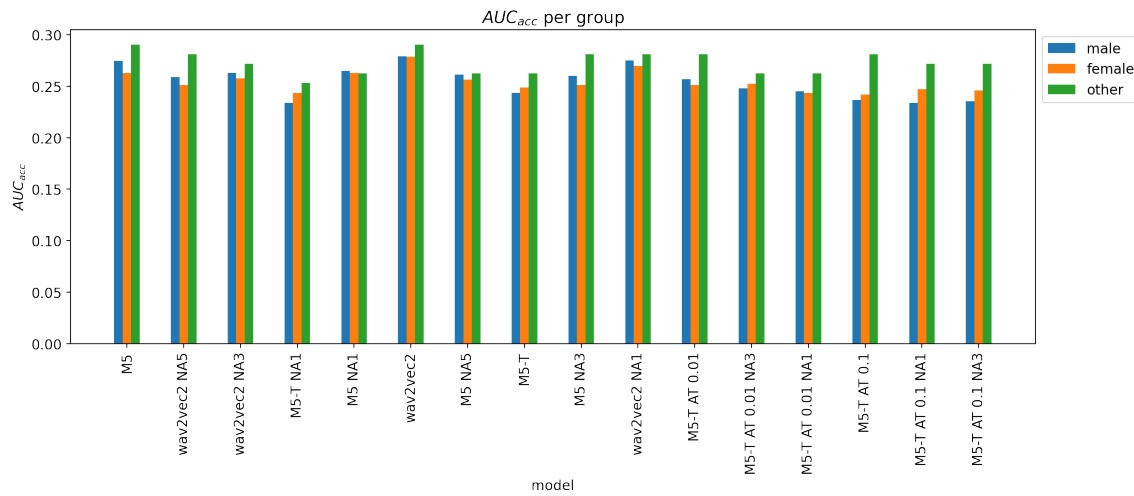
Fig. 8. A breakdown of gender groups overall AUC_{acc} per model.
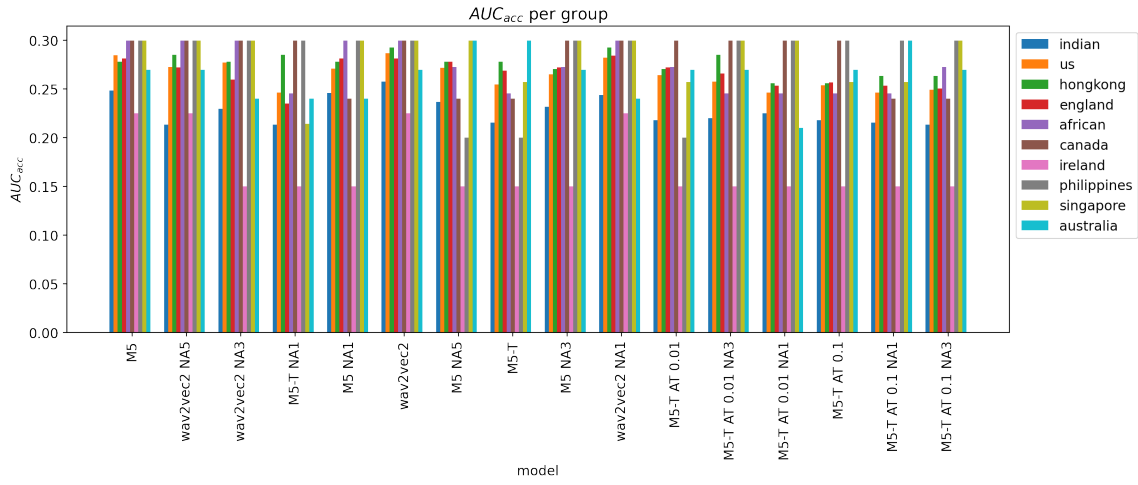
Fig. 9. A breakdown of accent groups overall AUC_{acc} per model.
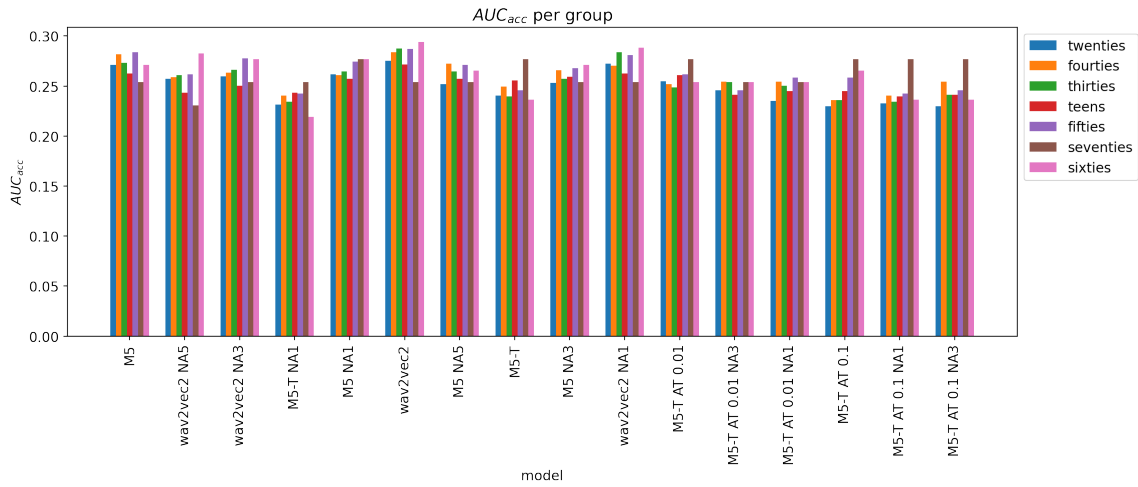
Fig. 10. A breakdown of age groups overall AUC_{acc} per model.
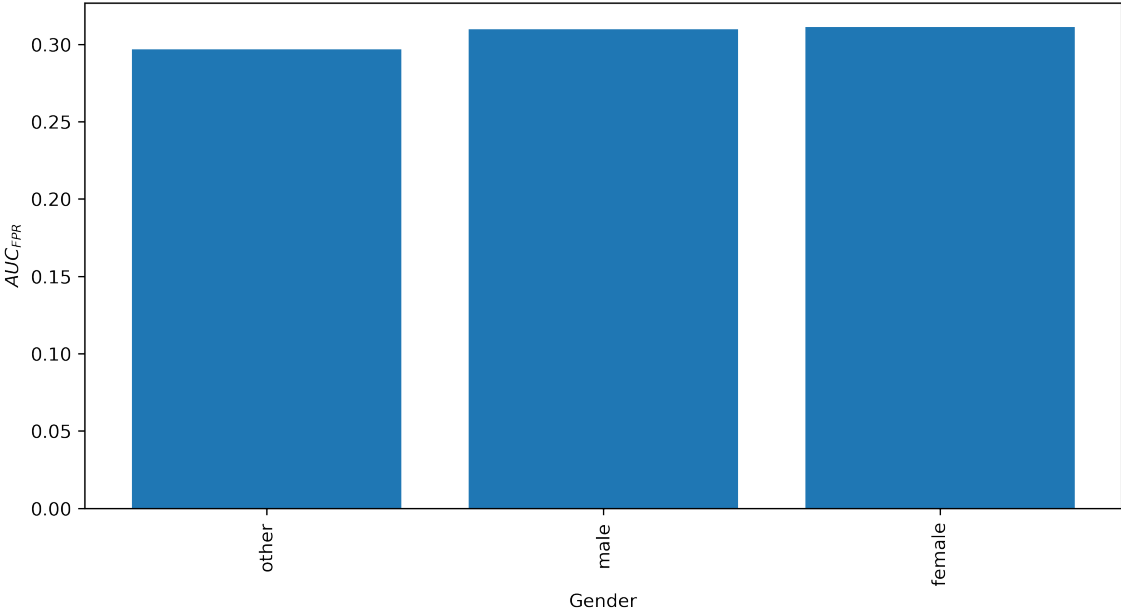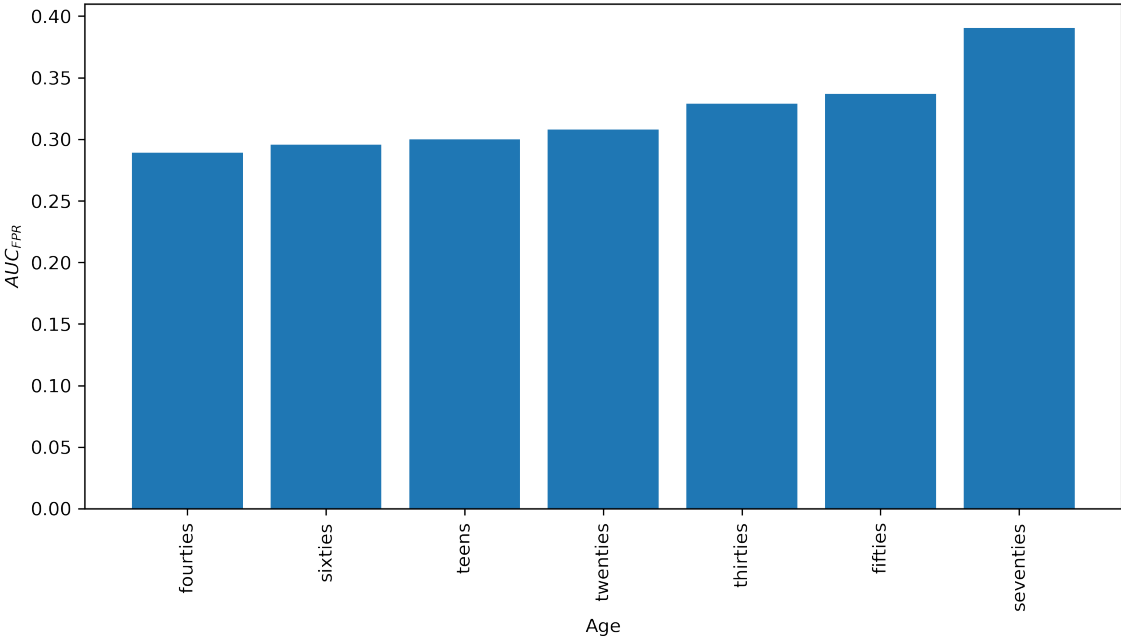
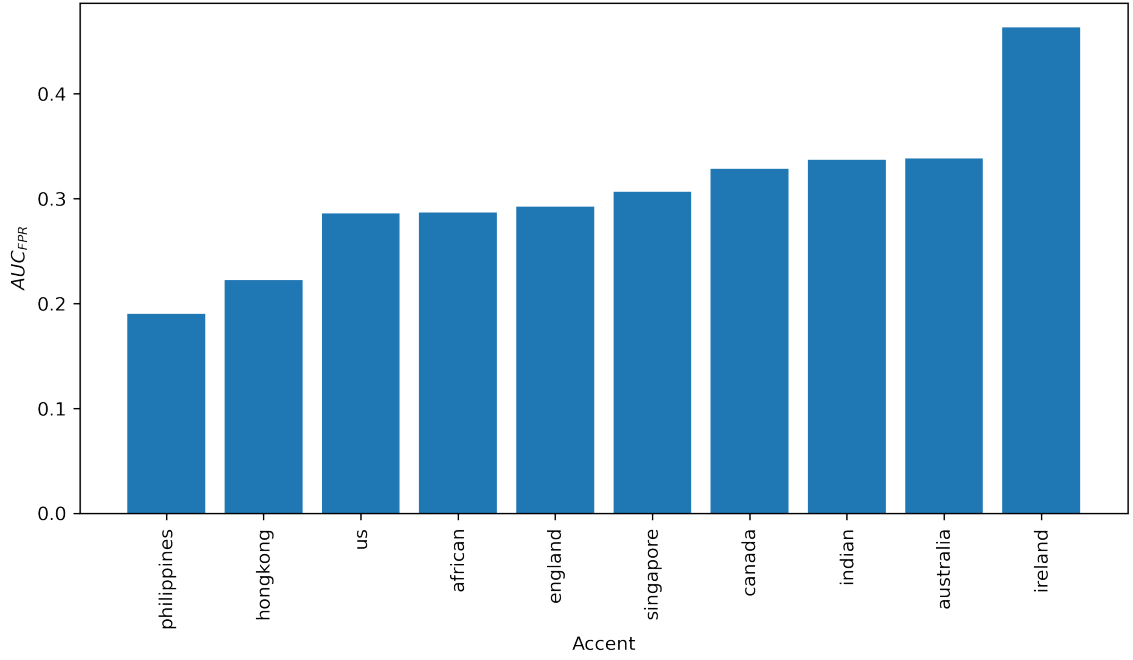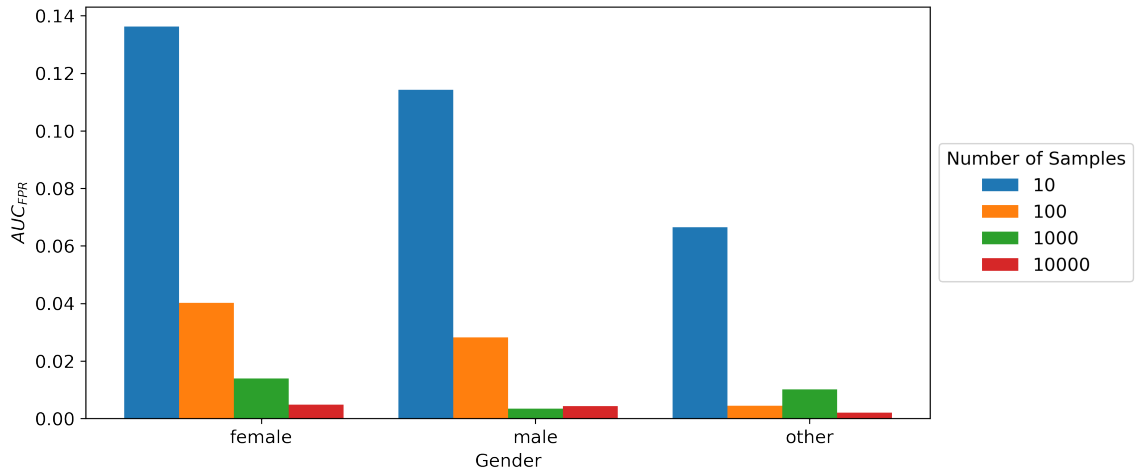Fig. 11. AUC_{FPR} of neural rejection for gender groups.

Fig. 12. AUC_{FPR} of neural rejection for age groups.

Fig. 13. AUC_{FPR} of neural rejection for accent groups.

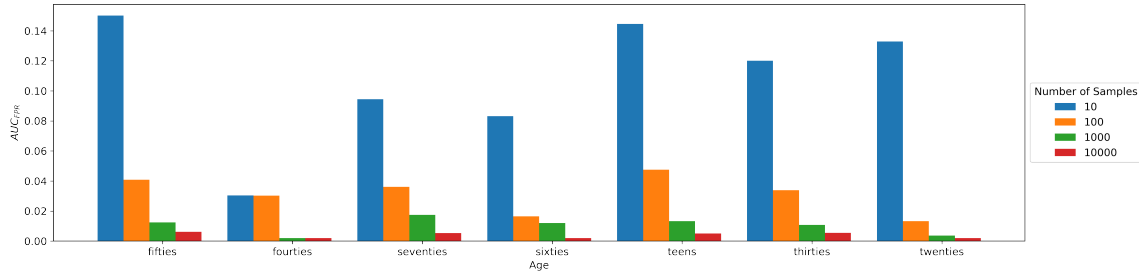Fig. 14. AUC_{FPR} of randomized smoothing rejection for gender groups.

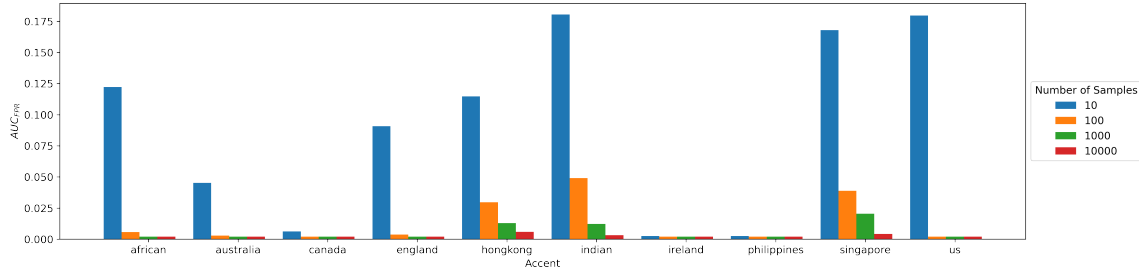Fig. 15. AUC_{FPR} of randomized smoothing rejection for age groups.

Fig. 16. AUC_{FPR} of randomized smoothing rejection for accent groups.