This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

FairSSD: Understanding Bias in Synthetic Speech Detectors

Amit Kumar Singh Yadav† Kratika Bhagtani† Davide Salvi‡ Paolo Bestagini‡ Edward J. Delp†

†Video and Image Processing Lab (VIPER), Purdue University, West Lafayette, Indiana, USA
‡Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milano, Italy

Abstract

Methods that can generate synthetic speech which is perceptually indistinguishable from speech recorded by a human speaker, are easily available. Several incidents report misuse of synthetic speech generated from these methods to commit fraud. To counter such misuse, many methods have been proposed to detect synthetic speech. Some of these detectors are more interpretable, can generalize to detect synthetic speech in the wild and are robust to noise. However, limited work has been done on understanding bias in these detectors. In this work, we examine bias in existing synthetic speech detectors to determine if they will unfairly target a particular gender, age and accent group. We also inspect whether these detectors will have a higher misclassification rate for bona fide speech from speech-impaired speakers w.r.t fluent speakers. Extensive experiments on 6 existing synthetic speech detectors using more than 0.9 million speech signals demonstrate that most detectors are gender, age and accent biased, and future work is needed to ensure fairness. To support future research, we release our evaluation dataset, models used in our study and source code at https://gitlab.com/viper-purdue/fairssd.

1. Introduction

Speech signals can be bona fide (*i.e.*, recorded from a human speaker) or synthetic (*i.e.*, generated from a computer) [1, 2]. With development in Generative Artificial Intelligence (AI) [3–5], methods to generate high-quality synthetic speech are easy available [6–9]. Many of these methods can mimic characteristics of an individual's voice, for example, its patterns, intonations, and pronunciations, which makes the synthesized speech perceptually indistinguishable from the recorded real one. Several recent incidents have reported misuse of such high-quality synthetic speech for spreading misinformation and being able to commit financial fraud [10–13].

To prevent the malicious use of synthetic speech, existing works have proposed several synthetic speech detectors [14–20]. Moreover, the multimedia forensics community has organized challenges and released datasets, such as ASVspoof2019 [21], to foster research and progress in this domain. Some of the released detectors focus on detecting synthetic speech in in-the-wild conditions (evaluation on speech synthesized from generation methods not seen before during training) [19, 22, 23]. Other methods aim to produce interpretable results [24–28]. Finally, other methods are claimed as robust to compression [15, 23, 29]. Despite the great efforts of multimedia forensic researchers, there is very limited work in understanding bias in these synthetic speech detectors. By bias, we refer to an action in which a detector unfairly targets a specific demographic group of individuals and falsely labels their bona fide speech as synthetic. Fig. 1 shows one such bias condition from our study in which the mean misclassification rate (False Positive Rate (FPR)) for all detectors is higher on bona fide speech from speech-impaired speakers than on bona fide speech from fluent speakers.

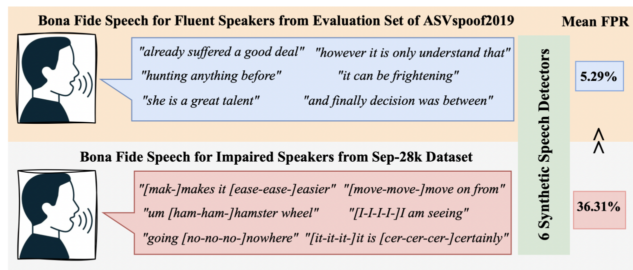

Figure 1. Mean False Positive Rate (FPR) of 6 synthetic speech detectors on bona fide speech from fluent and speech-impaired speakers.

Several existing works have examined bias in Automated Speaker Recognition (ASR) systems and deepfake face detectors [30–33]. For example, Pu *et al*. [34] observed that a deepfake face detector performs better for females as compared to males. Similar bias with respect to gender, age, speech impairment, race, and accent has also been shown in ASR systems [30]. However, limited attempts have been made in understanding bias in synthetic speech detectors. Ensuring fairness in synthetic speech detectors is important to prevent any misclassifications of bona fide speech from a particular ethnic and demographic group as synthetic, when these detectors are deployed on social platforms. As this can impact public opinion, it can lead to unintended and significant societal and political consequences, reducing trust on synthetic speech detectors and eroding reputation of the social platform using biased detectors.

In this work, we address this issue by thoroughly investigating and analyzing bias in synthetic speech detectors. We examine bias in six different approaches for detecting synthetic speech shown in Fig. 2 and described in Sec. 2.1. In our first 3 experiments, we examine whether synthetic speech detectors target a particular gender, age and accent group. We processed Mozilla Common Voice Corpora [35] and obtained approximately 0.9 million bona fide speech signals for our first 3 experiments. In our last experiment, we examine if synthetic speech detectors misclassify bona fide speech from people with speech impairments such as stuttering. We used bona fide speech from Sep-28K [36, 37] dataset for this study. Our results reveal hidden bias in existing synthetic speech detectors and we believe it will bring attention of forensics community to address these biases.

2. Related Work

In this section, we describe methods for synthetic speech detection and existing work on the fairness of forensics detectors.

2.1. Synthetic Speech Detection

An overview diagram of existing approaches for detecting synthetic speech is shown in Fig. 2. Based on the input, synthetic speech detection methods can be split into three categories.

The first and most conventional category of approaches obtains hand-crafted features such as Mel Frequency Cepstral Coefficients (MFCCs) [38], Constant Q Cepstral Coefficients (CQCCs) [39], bit-rate [40], and Linear Frequency Cepstral Coefficients (LFCCs) [41] from the speech signal to detect synthetic speech. Some methods process the obtained hand-crafted features using Gaussian Mixture Models (GMMs) [1, 42] and others use deep neural networks such as Bidirectional Long Short-Term Memory (Bi-LSTM) [43] and Residual Neural Network (ResNet) [44, 45].

The second category of approaches involves processing spectrogram images of a speech signal [17, 18, 23, 46, 47]. The spectrogram is a 2D representation that plots temporal variations in magnitude of different frequency components of the speech signal [48]. The frequency scale can be linear [46], logarithmic [18, 44] or it can be based on mel-filters [47, 48]. These methods then process the spectrogram using Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) [17, 46]. Some recent methods in this approach use self-supervised learning and train networks such

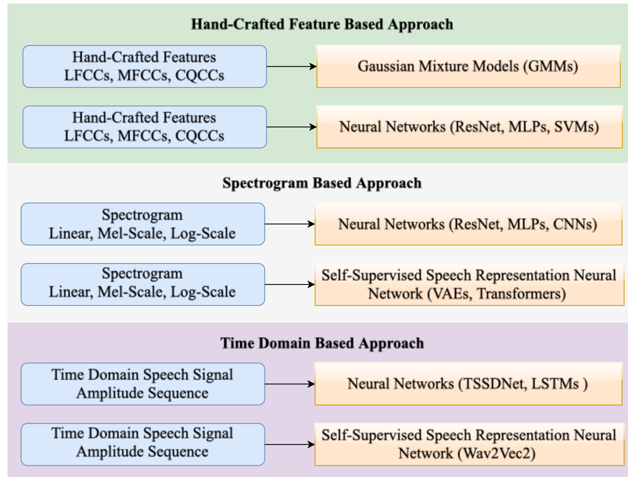

Figure 2. Existing Approaches for Synthetic Speech Detection.

as Variational Autoencoders (VAEs) [49] or a transformer neural network [50] on millions of real speech signals to first obtain a general speech representation network [51–53] and then fine-tune that network on limited training data for synthetic speech detection [23, 28, 47]. Such approaches are also effective in localizing the synthetic region and attributing the synthesizer used for generating it [54–57]

Methods in the third category process the time domain representation of the speech signal [14,19,58,59]. By time domain speech signal, we refer to 1D sequence that corresponds to temporal change in the amplitude of the speech signal. The classification network can either be a neural network trained from scratch *e.g.*, Time-Domain Synthetic Speech Detection Network (TSSDNet) and Long Short-Term Memory (LSTM) network as in [14, 19, 58, 59] or a pre-trained general speech representation neural network such as wav2vec 2.0 [16, 60, 61]. Most of the methods use AudioSet [62] or Libri-Speech dataset [63] for any pre-training to obtain general speech representation neural network [51–53, 60]. For fine-tuning or from-scratch training and evaluation almost all the methods use ASVspoof2019 [21] dataset.

2.2. Fairness of Forensic Detectors

Several existing works examine and improve generalization capability [15, 16, 19, 22], robustness to compression [23, 29], and explainability [24, 28] of synthetic speech detectors. Some work has been done in understanding bias in ASR systems [30–32]. Many works have explored bias in face recognition systems [64–66]. These have been extended to evaluate bias in deepfake face detectors [33, 34, 67–70]. Trinh *et al.* [70] examined bias in three existing deepfake face detectors and observed a disparity in performance of more than 10 percentage points between different demographic groups. Hazirbas *et al.* [66] had similar observations.

Pu *et al*. [34] evaluated a popular deepfake face detector and observed that it performs better for female gender. Xu *et al*. [33] annotated a deepfake detection dataset and performed a thorough bias evaluation of deepfake face detectors. Nadimpalli *et al*. [71] observed gender bias in deepfake face detectors and worked on reducing it by training on a more gender-balanced dataset. However, such an approach requires time-consuming data annotation. To counter this, Ju *et al*. [72] presented a loss function for improving fairness of deepfake detectors. This loss function can work in both scenarios, *i.e*., when demographic annotations are either available or absent during training. While many initial works have analyzed and showed evidence of biases in deepfake face detectors, reducing those biases is still an open research area [67]. In this work, similar to previous work examining biases in deepfake face detectors, we attempt to understand biases in synthetic speech detectors.

3. Proposed Study

In this section, we describe the synthetic speech detectors examined in this work, the dataset used for training them, the evaluation datasets and metrics used in our bias study.

3.1. Detectors Used in Our Study

An overview of different categories of synthetic speech detection methods is provided in Fig. 2 and Sec. 2.1. To run a comprehensive study, we include methods from each category shown in Fig. 2. In total, we examine 6 different synthetic speech detection methods. Out of these 6 methods, we trained 4 of them from scratch to obtain detectors for our study. For the remaining 2, we perform our study directly on the model and weights released by the authors. More details about each method and its training setup are presented below.

LFCC-GMMs [1, 42] : We use this method as a representative of methods that process hand-crafted features using GMMs. It is the best hand-crafted baseline provided in the ASVspoof2019 [21] challenge. In this method, LFCCs [42] are obtained from the speech signal. After obtaining the LFCCs features, for each class (bona fide speech class and synthetic speech class), we fit a separate GMM for each class. Each GMM estimates the probability distribution of the features for that class. During evaluation, we obtain LFCCs from a given speech signal and use GMMs to obtain the probability of it belonging to each class. The speech signal is labeled as the most probable class. We performed an ablation study to determine the best set of hyper-parameters for obtaining the hand-crafted features. More details about the ablation study and training can be found in the supplement material. The model used in this study processes 30ms windows with a hop size of 15ms and frequency components up to 4KHz to obtain LFCCs features. GMM for each class has 512 Gaussian mixture components.

MFCC-ResNet [44] : We use this method as a representative of the methods that obtain hand-crafted features and process them using a neural network. For this method, we trained and obtained our own model weights. This method uses Mel-Frequency Cepstrum Coefficients (MFCC) [38], which were also used in baselines provided in the ASVspoof2019 challenge [1]. The MFCC are obtained using the Short Time Fourier Transform (STFT) of the speech signal, mel-spectrum filters and Discrete Cosine Transform (DCT) [44]. For our experiments, we select the first 24 MFCC coefficients. Besides MFCC, we also obtain the first and second derivatives of MFCC. Overall, we use features of dimension 72. These hand-crafted features are processed by a ResNet [45] for synthetic speech detection.

Spec-ResNet [44] : We use this method as a representative for methods that process spectrograms using a neural network. For this method, we trained and obtained our own model weights. This method obtains the spectrogram by taking STFT of the input speech and squaring its magnitude. The scale is changed to a logarithmic scale. A ResNet network [45] processes the logarithmic-scaled spectrogram for synthetic speech detection.

PS3DT [23] : This method is representative of methods that process spectrogram representation of the speech signal using a self-supervised pre-trained network. We implemented and trained this method. We selected this method as it performs better than others which use a similar approach [47] and process spectrograms. It also has better in-the-wild performance and is more robust to compression [23]. It uses mel-scale [48] spectrogram. The network is pre-trained on the Audioset Dataset [62] using self-supervised learning. PS3DT divides a mel-spectrogram into patches and obtains patch representations using a transformer neural network. During pre-training, some patches are masked and reconstructed [73]. Later, the pre-trained network is fine-tuned on ASVspoof2019 training set [21]. During fine-tuning, the patch representations are rearranged such that representations corresponding to the same temporal location are aligned together. A Multi Layer Perceptron Network (MLP) process these representations to label speech as bona fide or synthetic.

TSSDNet [14] : This method is representative of an end-to-end method that processes time-domain speech signals. We used the weights released by authors in [14]. We used this method as it is shown in [14, 74] that it performs better than other similar time-domain methods such as Raw PC-DARTS [75] and is also found robust to compression noise [29]. In TSSDNet, two kinds of networks are proposed, namely, ResNet style and Inception style networks, which are used to process speech amplitude [14]. In [14], the ResNet version performs better than the Inception version, and therefore we use the ResNet version in our experiments.

Wav2Vec2-AASIST [16]: This method is representative of methods that process time-domain speech signals using a large self-supervised neural network. We used the weights

Table 1. Details of speech signals in bona fide class of each evaluation set used for gender, accent and age bias studies.

Study	Set Name	Accent	Age Group	Gender	Samples
Gender	$D_{US-20s-M}$	US English	20s	Male	31,000
	$D_{US-20s-F}$	US English	20s	Female	31,000
	$D_{US-30s-M}$	US English	30s	Male	15,000
	$D_{US-30s-F}$	US English	30s	Female	15,000
	$D_{US-60s-M}$	US English	60s	Male	16,000
	$D_{US-60s-F}$	US English	60s	Female	16,000
Age	$D_{US-ts-F}$	US English	teens	Female	8,900
	$D_{US-20s-F}$	US English	20s	Female	8,900
	$D_{US-30s-F}$	US English	30s	Female	8,900
	$D_{US-40s-F}$	US English	40s	Female	8,900
	$D_{US-50s-F}$	US English	50s	Female	8,900
	$D_{US-60s-F}$	US English	60s	Female	8,900
	$D_{US-ts-M}$	US English	teens	Male	8,900
	$D_{US-20s-M}$	US English	20s	Male	8,900
	$D_{US-30s-M}$	US English	30s	Male	8,900
	$D_{US-40s-M}$	US English	40s	Male	8,900
	$D_{US-50s-M}$	US English	50s	Male	8,900
	$D_{US-60s-M}$	US English	60s	Male	8,900
Accent	$D_{US-20s-F}$	US English	20s	Female	4,900
	$D_{SA-20s-F}$	South Asian	20s	Female	4,900
	$D_{CN-20s-F}$	Canadian	20s	Female	4,900
	$D_{UK-20s-F}$	British	20s	Female	4,900
	$D_{AU-20s-F}$	Australian	20s	Female	4,900
	$D_{US-20s-M}$	US English	20s	Male	8,100
	$D_{SA-20s-M}$	South Asian	20s	Male	8,100
	$D_{CN-20s-M}$	Canadian	20s	Male	8,100
	$D_{UK-20s-M}$	British	20s	Male	8,100
	$D_{AU-20s-M}$	Australian	20s	Male	8,100

provided by the authors in [16]. We use this method as it has nearly perfect performance on ASVspoof2019 [21] and has better generalization performance than several existing methods on the ASVspoof2021 dataset [16]. This method processes the time domain speech signal using a pre-trained wav2vec2 network [60] and combines it with spectral features using the Audio Anti-Spoofing using Integrated Spectro-Temporal (AASIST) graph attention network [74]. Wav2vec2 has more than 300 million parameters and is trained on a large audio dataset using self-supervision. It consists of a CNN and a transfomer network [16, 50].

3.2. Datasets Used in Our Study

In this section, we first describe the dataset used for training the detectors and evaluating their detection performance. We later discuss our evaluation dataset prepared for examining the age, gender and accent bias in each detector. We also discuss the dataset used for the bias study on speech with stuttering impairment.

3.2.1 Detection Training and Evaluation Dataset

Following most existing work on synthetic speech detection [14–16], we use Logical Access (LA) part of the ASVSpoof2019 Dataset [21] for training. Each of the existing methods is trained, validated and evaluated on the official training set (D_{tr}), development set (D_{dev}) and evaluation set (D_{eval}), respectively. There are 25,380 speech signals in D_{tr}, 24,844 in D_{dev}, and 71,237 speech signals in D_{eval} out of which 2580, 2548, and 7355, respectively, are bona fide speech signals. The synthetic speech signal in D_{tr} and D_{dev} are generated using 6 different synthetic speech generators. The synthetic speech signals in D_{eval} are generated from 13 different speech generators, 2 generators are the same as the ones used in D_{tr} and D_{dev} sets, while 11 generators are unknown generators not present in D_{tr} and D_{dev} sets. Hence, the evaluation set has a majority of synthetic speech signals generated from unknown speech synthesizers not seen during training. Speakers recorded for bona fide speech do not overlap among any sets and the sampling rate is 16 kHz. In all our bias studies, we kept the same sampling rate during evaluation.

3.2.2 Evaluation Datasets for Bias Study

In our study, we want to examine if existing synthetic speech detectors will unfairly target bona fide speech from a particular group and misclassify it as synthetic. To facilitate our first 3 studies which examine gender, age and accent bias, we need bona fide speech signals with demographic annotations such as gender, age, and accent. We use bona fide speech samples from multi-language speech corpus Mozilla Common Voice Corpus 16.1 [35]. We use only the validated and the English subset of the Mozilla CVC dataset [35] that has approximately 1.78 million speech signals, in total constituting 2,600 hours of bona fide speech. Many of these speech signals do not have all the required annotations for our study. We pre-processed the dataset and obtained approximately 0.9 million speech signals having required annotations. From these, we constructed 28 evaluation sets for performing our bias study. All 28 evaluation sets consist of speech signals from bonafide class as well as synthetic class.

First, we describe the details of speech from bonafide class, which comes from Mozilla CVC dataset [35]. The number of bona fide signals in each evaluation set is reported in Tab. 1. We selected the numbers based on the minimum number of samples present in one group, which is explained next. For example, in our gender bias study with speakers in age group 20s that have US accent, we have two sets, namely $D_{US-20s-M}$ and $D_{US-20s-F}$ as shown in Tab. 1. The biggest set with demographic: female, US accent and in age group 20s has 31.5K bona fide speech signals and similarly the biggest set with demographic: male, US accent and in age group 20s has 110K bona fide speech signals. We randomly sample 31K speech signals from each of the male and female sets because this is the approximate minimum of both numbers. For consistency, we need to select equal number of speech signals from male and female genders. We repeated this random sampling 5 times, leading to 5 versions of each set mentioned in Tab. 1. Therefore, each experiment is run 5 times and the mean and standard deviation for each performance metric is provided in all of our bias studies.

Next, we describe the details of the synthetic class of all 28 evaluation sets. The synthetic speech class is same in each set and it consists of all the synthetic speech signals from the D_{eval} set of ASVspoof2019 Dataset. This is done so that any difference in the performance of a detector on two different sets can directly be attributed to the difference in its performance on demographic of bonafide class in the sets. We created 6 sets for our gender bias study. We fixed accent to US English and evaluated for 3 different ages (see Tab. 1). For each age, we created two sets, one with bona fide speech from only male speakers and other with only female speakers. Similarly, we created 10 sets for understanding accent bias and 12 sets for examining age bias. More details about the dataset and pre-processing are provided in the supplement.

For examining bias on stuttering speech, we use bona fide speech from SEP-28k dataset [36, 37] and synthetic speech from D_{eval} set of ASVspoof2019 dataset. It contains approximately 28K bona fide speech signals. The stuttering speech can have prolongation: elongated syllable (*e.g.*, M[mmm]ommy), gasps for air or stuttered pauses, repeated syllables (*e.g.*, I [pr-pr-pr-]prepared dinner), word or phrase repetition (*e.g.*, I made [made] dinner), and filler words use to cope with stutter (*e.g.*, "um" or "uh"). We pre-processed the bona fide speech to remove speech samples with poor audio quality, music in background, and signals with just silence or background noise. After pre-processing, we obtained 21,855 bona fide stuttering speech signals.

3.3. Evaluation Metrics

We use two metrics for evaluation. First, we use Equal Error Rate (EER) for measuring performance of each detector on ASVspoof2019 dataset. EER is the performance metric used in ASVspoof2019 challenge [1]. For calculating EER, a decision threshold is determined that balances and makes both FPR and False Negative Rate (FNR) equal. The EER is same as the FPR or FNR obtained at that threshold. We obtain EER on ASVspoof2019 evaluation set (D_{eval}) as a performance measure of an individual synthetic speech detector. Lower the EER, the better the performance of the detector. EER is also used in existing speech recognition work for bias study [30, 31]. Difference in EER values by a detector on *e.g.* male and female sets is a measure of the gender bias. For example, in our gender bias study we use $D_{US-20s-M}$ and $D_{US-20s-F}$ sets shown in Tab. 1. The synthetic speech class of each set is same and both sets have same number of bona fide speech. As shown in Tab. 1, the age and accent for samples in bona fide class is also the same. Hence, the two sets only differ in terms of bona fide speech gender. Therefore, the difference in EER performance by a detector on $D_{US-20s-M}$ and $D_{US-20s-F}$ will imply that the detector targets bona fide speech from one group unfairly and is biased w.r.t gender. In all our bias studies such as gender bias study (Tab. 3), for each evaluation set, we report $\Delta EER := EER - minEER$ as a bias measure, where $minEER$ is minimum EER of a detector obtained for a particular demographic group. For example, in our gender bias study for speakers in age group 20s, we use $D_{US-20s-M}$ and $D_{US-20s-F}$ sets from Tab. 1. We obtain EER of a particular detector on both sets. The minimum of the two obtained EERs is $minEER$ for that detector and used to report ΔEER on $D_{US-20s-M}$ and $D_{US-20s-F}$ sets. Note: ΔEER will always be positive. It will be zero for either one of $D_{US-20s-M}$ or $D_{US-20s-F}$ as one of the EER values will be same as $minEER$. The higher ΔEER for the other set, the more biased the detector is for that set (gender) and bona fide speech from that group has been more unfairly mislabeled as synthetic. The absolute values of EER and more details about calculating ΔEER are provided in the supplement material.

The EER provides an idea of a detector performance under the assumption that the distribution of the data under analysis is coherent with the distribution of the evaluation set used for the tests. This may not be true, particularly if the detector is deployed on some platform for real-world use-case. Hence, we also report the FPR of each detector on sets shown in Tab. 1. FPR measures the percentage of bona fide speech signals falsely labeled as synthetic. All existing methods provide the probability of a speech signal being synthetic. Therefore, the decision to label a speech signal as synthetic requires comparing the obtained probability from each method with a threshold. We obtain the threshold for calculating FPR on each set shown in Tab. 1 using an independent dataset. We use the evaluation set (D_{eval}) of ASVspoof2019 to obtain the threshold. Speech signals in D_{eval} set are generated from unknown speech generators and bona fide speakers do not overlap with the training set, making it a reasonable dataset for obtaining the threshold.

We obtain three FPRs representing three different constraints. The threshold for FPR_1 is where the detector has equal FPR and FNR on D_{eval} set of ASVspoof2019 dataset. The threshold for FPR_2 is where the detector has 0.08FPR on D_{eval} set of ASVspoof2019 dataset. The threshold for FPR_3 is where the detector has 0.08FNR on D_{eval} set of ASVspoof2019 dataset. Similar to the difference in EER values, the difference inFPR values for an individual detector (*e.g.*, male and female subsets) is a measure of bias (*e.g.*, gender bias). Hence, in all bias studies, similar to EER, we report ΔFPR_1, ΔFPR_2, and ΔFPR_3. Each of them will be a positive value. Higher the ΔFPR, the more biased the detector is for that particular evaluation set (demographic). The absolute values and more details about ΔFPRs are provided in the supplement material.

Table 2. EER (in %) performance of detectors on ASVspoof2019.

Detector Number (DN)	Name	Type	Parameters	D_{dev}	D_{eval}
D01 [14]	TSSDNet	Time-domain	0.35M	0.74	1.62
D02 [16]	Wav2Vec2-AASIST	Time-domain	317M	0.02	0.23
D03 [44]	Spec-ResNet	Log-spectrogram	0.32M	0.71	10.10
D04 [23]	PS3DT	Mel-spectrogram	95M	2.82	4.54
D05 [1]	LFCC-GMMs	Hand-crafted	0.1M	0.04	3.67
D06 [44]	MFCC-ResNet	Hand-crafted	0.26M	6.52	11.58
			Mean	1.81	5.29

4. Experiments and Results

In this section, we discuss the performance of all detectors on ASVspoof2019 Dataset and experimental results from our bias studies. We used D_{dev} and D_{eval} sets of ASVspoof2019 Dataset for measuring detection performance. For bias studies, we used the evaluation dataset described in Sec. 3.2.

4.1. Experiment 1: Detection Performance

We evaluate all detectors on the development set (D_{dev}) and the evaluation set (D_{eval}) of the ASVspoof2019 Dataset. The goal of this experiment is to check that the selected detectors have been correctly trained and work well for synthetic speech detection.

The results of this evaluation are shown in Tab. 2. The performance on D_{dev} is representative of closed-set performance as synthetic speech is generated from the same synthesizers as used in the training set. The performance on D_{eval} is representative of the generalization capability of an individual detector to detect synthetic speech from unknown speech generators. As 11 out of 13 speech generators used in D_{eval} set are unknown and speech signals from them were not used during training. All methods except MFCC-ResNet have almost perfect performance on D_{dev} set (EER 0%). The EER on D_{eval} set, as expected, is higher than on D_{dev} set for each method. For both time domain and spectrogram-based approaches, the detectors which use self-supervised networks have substantially better performance on D_{eval} set, however, that comes at a computational cost: large training time and substantially higher model parameters. For 3 methods (in Tab. 2), our implementation has different EER on D_{eval} set than reported in their original work. The LFCC-GMMs implemented by us has EER of 3.67% on D_{eval} set as shown in Tab. 2, while the EER reported in [1] is 8.09%. Implementation in [1] was done in MATLAB, but we used Python and scikit-learn for our implementation. Our ablation study reported in supplement material shows that the hyper-parameters are not the reason for this change as we also got better performance using same hyperparameters reported in [1]. For Spec-ResNet and MFCC-ResNet, the EER from our models weights (refer Tab. 2) on D_{eval} set are 0.42 and 2.25 percentage points, respectively, higher than EER reported in [44]. We use same hyper-parameters but an updated version of the feature extraction package as compared to the version used by [44], as discussed in the supplement material.

Table 3. Gender Bias Study for all detectors. Values are in %.

DN	Metric	$D_{US-20s-M}$	$D_{US-20s-F}$	$D_{US-30s-M}$	$D_{US-30s-F}$	$D_{US-60s-M}$	$D_{US-60s-F}$
D01	ΔFPR_1	1.47 ± 0.087	0.00 ± 0.012	0.92 ± 0.034	0.00 ± 0.018	**6.56 ± 0.076**	**0.00 ± 0.105**
	ΔFPR_2	**0.36 ± 0.009**	**0.00 ± 0.006**	0.10 ± 0.012	0.00 ± 0.005	0.00 ± 0.006	0.03 ± 0.009
	ΔFPR_3	0.80 ± 0.162	0.00 ± 0.053	6.01 ± 0.282	0.00 ± 0.101	**26.12 ± 0.248**	**0.00 ± 0.348**
	ΔEER	1.45 ± 0.053	0.00 ± 0.026	0.00 ± 0.171	1.21 ± 0.024	**14.84 ± 0.062**	**0.00 ± 0.083**
D02	ΔFPR_1	0.00 ± 0.057	2.09 ± 0.047	0.00 ± 0.313	2.40 ± 0.051	**11.40 ± 0.208**	**0.00 ± 0.288**
	ΔFPR_2	0.00 ± 0.132	0.61 ± 0.026	**0.00 ± 0.381**	**6.73 ± 0.048**	0.38 ± 0.101	0.00 ± 0.136
	ΔFPR_3	0.32 ± 0.112	0.00 ± 0.020	**0.93 ± 0.095**	**0.00 ± 0.041**	0.78 ± 0.019	0.00 ± 0.013
	ΔEER	0.20 ± 0.039	0.00 ± 0.013	0.71 ± 0.080	0.00 ± 0.015	**1.16 ± 0.038**	**0.00 ± 0.050**
D03	ΔFPR_1	**0.15 ± 0.030**	**0.00 ± 0.005**	0.05 ± 0.009	0.00 ± 0.005	0.12 ± 0.010	0.00 ± 0.013
	ΔFPR_2	**0.14 ± 0.022**	**0.00 ± 0.009**	0.11 ± 0.059	0.00 ± 0.008	0.13 ± 0.018	0.00 ± 0.025
	ΔFPR_3	**0.08 ± 0.016**	**0.00 ± 0.005**	0.02 ± 0.012	0.00 ± 0.008	0.08 ± 0.017	0.00 ± 0.025
	ΔEER	**2.96 ± 0.035**	**0.00 ± 0.012**	1.02 ± 0.053	0.00 ± 0.006	2.02 ± 0.020	0.00 ± 0.026
D04	ΔFPR_1	22.48 ± 0.001	0.00 ± 0.034	**39.87 ± 0.280**	**0.00 ± 0.032**	11.61 ± 0.135	0.00 ± 0.177
	ΔFPR_2	22.32 ± 0.195	0.00 ± 0.052	**39.52 ± 0.253**	**0.00 ± 0.072**	11.17 ± 0.157	0.00 ± 0.207
	ΔFPR_3	22.80 ± 0.150	0.00 ± 0.023	**40.34 ± 0.246**	**0.00 ± 0.071**	11.78 ± 0.173	0.00 ± 0.229
	ΔEER	4.94 ± 0.089	0.00 ± 0.028	**14.80 ± 0.238**	**0.00 ± 0.035**	0.97 ± 0.049	0.00 ± 0.057
D05	ΔFPR_1	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000
	ΔFPR_2	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000
	ΔFPR_3	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000
	ΔEER	2.04 ± 0.072	0.00 ± 0.008	0.70 ± 0.115	0.00 ± 0.031	**0.00 ± 0.042**	**2.56 ± 0.049**
D06	ΔFPR_1	0.00 ± 0.099	0.41 ± 0.028	7.39 ± 0.178	0.00 ± 0.092	**0.00 ± 0.178**	**8.58 ± 0.190**
	ΔFPR_2	0.12 ± 0.171	0.00 ± 0.035	**10.74 ± 0.238**	**0.00 ± 0.029**	0.00 ± 0.158	5.74 ± 0.205
	ΔFPR_3	0.00 ± 0.102	0.90 ± 0.015	4.91 ± 0.170	0.00 ± 0.055	**0.00 ± 0.109**	**9.04 ± 0.147**
	ΔEER	1.39 ± 0.113	0.00 ± 0.028	**6.62 ± 0.139**	**0.00 ± 0.046**	1.10 ± 0.092	0.00 ± 0.097

4.2. Experiment 2: Studying Bias on Gender

In this experiment, we determine if a detector misclassifies bona fide speech more from one particular gender. The results for this experiment are shown in Tab. 3.

We fixed the accent to US English and studied gender bias on speaker in three different age groups, namely, speakers in their 20s, 30s, and 60s. For each age group, we have two sets. For example, $D_{US-20s-M}$ and $D_{US-20s-F}$ are two sets for speakers in age group 20s. We obtain ΔEER, ΔFPR_1, ΔFPR_2, and ΔFPR_3 as a measure of bias in different scenarios. For example, FPRs are obtained using threshold from an independent dataset. However, EER assumes that ground truth labels are known and it is the rate where FPR and FNR are equal (see Sec. 3.3).

In Tab. 3, we highlight most gender biased age group for each detector. For example, Detector Number (DN) D01 *i.e.*, TSSDNet has 26.12 percentage points higher FPR_3 for male speakers than for female speakers in 60s age group. Similar bias is also evident in ΔEER value of this group. Similar results are also obtained for male and female speakers in age group 20s, where D01 has 1.45 percentage points higher EER on male speakers than on female speakers and all FPRs are higher on male speakers. For speakers in 30s, the FPR is similarly higher for male speakers. Overall, these observations suggest that D01 misclassifies bona fide speech from male speakers more than that from female speakers. Similar observations can also be noted for D04 *i.e.*, PS3DT which has approximately 50 percentage points of more misclassifications for bona fide speech from male

Table 4. Age Bias Study for Male Speakers. Values are in %.

DN	Metric	$D_{US-ts-M}$	$D_{US-20s-M}$	$D_{US-30s-M}$	$D_{US-40s-M}$	$D_{US-50s-M}$	$D_{US-60s-M}$
D01	ΔFPR_1	0.00 ± 0.109	1.03 ± 0.093	1.32 ± 0.179	0.54 ± 0.092	0.54 ± 0.092	**1.36 ± 0.131**
	ΔFPR_2	0.03 ± 0.026	0.04 ± 0.030	**0.07 ± 0.017**	0.03 ± 0.019	0.01 ± 0.020	0.00 ± 0.021
	ΔFPR_3	2.51 ± 0.261	6.11 ± 0.438	6.62 ± 0.375	0.00 ± 0.282	2.21 ± 0.231	**15.34 ± 0.361**
	ΔEER	2.57 ± 0.368	4.50 ± 0.473	1.22 ± 0.368	0.00 ± 0.475	2.37 ± 0.359	**15.89 ± 0.360**
D02	ΔFPR_1	**11.88 ± 0.395**	5.30 ± 0.323	0.00 ± 0.249	4.32 ± 0.354	5.31 ± 0.278	6.90 ± 0.384
	ΔFPR_2	8.18 ± 0.463	6.80 ± 0.457	0.00 ± 0.455	1.23 ± 0.440	3.54 ± 0.370	**9.25 ± 0.336**
	ΔFPR_3	**2.62 ± 0.171**	1.16 ± 0.147	0.88 ± 0.198	2.56 ± 0.196	1.76 ± 0.152	0.00 ± 0.103
	ΔEER	**2.10 ± 0.106**	0.83 ± 0.129	0.71 ± 0.154	1.93 ± 0.190	1.40 ± 0.109	0.00 ± 0.136
D03	ΔFPR_1	**0.45 ± 0.062**	0.30 ± 0.088	0.29 ± 0.075	0.00 ± 0.085	0.36 ± 0.065	0.43 ± 0.066
	ΔFPR_2	0.55 ± 0.068	0.36 ± 0.107	0.34 ± 0.086	0.00 ± 0.087	0.42 ± 0.069	**0.57 ± 0.066**
	ΔFPR_3	**0.22 ± 0.049**	0.12 ± 0.051	0.10 ± 0.057	0.00 ± 0.060	0.16 ± 0.043	0.22 ± 0.046
	ΔEER	2.27 ± 0.160	2.02 ± 0.141	0.34 ± 0.178	0.00 ± 0.172	0.98 ± 0.132	**3.19 ± 0.132**
D04	ΔFPR_1	0.00 ± 0.773	5.52 ± 0.776	**12.11 ± 0.622**	10.24 ± 0.736	11.21 ± 0.596	6.70 ± 0.576
	ΔFPR_2	0.00 ± 0.599	5.58 ± 0.588	11.87 ± 0.548	10.49 ± 0.559	**11.42 ± 0.432**	7.15 ± 0.621
	ΔFPR_3	0.00 ± 0.413	5.28 ± 0.841	**12.34 ± 0.416**	10.99 ± 0.356	11.16 ± 0.338	6.91 ± 0.355
	ΔEER	0.24 ± 0.188	1.26 ± 0.132	**7.14 ± 0.200**	1.43 ± 0.253	3.04 ± 0.110	0.00 ± 0.103
D05	ΔFPR_1	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000
	ΔFPR_2	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000
	ΔFPR_3	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000
	ΔEER	0.00 ± 0.165	1.43 ± 0.213	**3.19 ± 0.131**	1.07 ± 0.169	1.89 ± 0.138	3.07 ± 0.165
D06	ΔFPR_1	15.27 ± 0.415	13.29 ± 0.434	**17.97 ± 0.509**	12.39 ± 0.406	11.41 ± 0.338	0.00 ± 0.446
	ΔFPR_2	16.23 ± 0.455	14.48 ± 0.472	**19.70 ± 0.402**	13.13 ± 0.684	11.30 ± 0.425	0.00 ± 0.441
	ΔFPR_3	12.22 ± 0.327	10.64 ± 0.323	**13.71 ± 0.286**	9.89 ± 0.425	9.05 ± 0.293	0.00 ± 0.328
	ΔEER	8.01 ± 0.185	7.58 ± 0.297	**10.33 ± 0.162**	7.33 ± 0.218	4.59 ± 0.223	0.00 ± 0.196

Table 5. Age Bias Study for Female Speakers. Values are in %.

DN	Metric	$D_{US-ts-F}$	$D_{US-20s-F}$	$D_{US-30s-F}$	$D_{US-40s-F}$	$D_{US-50s-F}$	$D_{US-60s-F}$
D01	ΔFPR_1	**6.54 ± 0.356**	4.88 ± 0.421	5.78 ± 0.350	5.75 ± 0.346	3.97 ± 0.351	0.00 ± 0.488
	ΔFPR_2	0.30 ± 0.075	0.00 ± 0.105	0.27 ± 0.076	0.19 ± 0.074	0.22 ± 0.076	**0.32 ± 0.076**
	ΔFPR_3	**19.83 ± 0.238**	18.58 ± 0.340	14.16 ± 0.252	0.00 ± 0.086	10.51 ± 0.304	2.56 ± 0.252
	ΔEER	5.92 ± 0.230	**6.82 ± 0.215**	6.09 ± 0.124	0.00 ± 0.049	4.40 ± 0.126	4.73 ± 0.077
D02	ΔFPR_1	**18.48 ± 0.459**	12.16 ± 0.538	6.42 ± 0.431	13.06 ± 0.325	5.87 ± 0.389	0.00 ± 0.458
	ΔFPR_2	3.71 ± 0.286	2.77 ± 0.258	1.29 ± 0.347	5.41 ± 0.225	0.00 ± 0.317	**4.04 ± 0.305**
	ΔFPR_3	**12.15 ± 0.233**	1.71 ± 0.185	0.75 ± 0.078	1.24 ± 0.044	2.07 ± 0.124	0.00 ± 0.058
	ΔEER	**9.61 ± 0.168**	1.70 ± 0.128	0.96 ± 0.138	1.28 ± 0.074	2.05 ± 0.094	0.00 ± 0.104
D03	ΔFPR_1	**0.80 ± 0.073**	0.63 ± 0.110	0.67 ± 0.075	0.69 ± 0.063	0.00 ± 0.089	0.78 ± 0.077
	ΔFPR_2	0.91 ± 0.056	0.73 ± 0.046	0.70 ± 0.040	0.82 ± 0.031	0.00 ± 0.044	**0.96 ± 0.062**
	ΔFPR_3	**0.35 ± 0.072**	0.26 ± 0.077	0.27 ± 0.076	0.30 ± 0.071	0.00 ± 0.100	0.33 ± 0.076
	ΔEER	0.00 ± 0.097	1.23 ± 0.093	1.58 ± 0.120	2.61 ± 0.071	0.21 ± 0.158	**3.42 ± 0.102**
D04	ΔFPR_1	20.47 ± 0.377	11.28 ± 0.432	0.00 ± 0.441	13.79 ± 0.312	1.22 ± 0.358	**22.88 ± 0.594**
	ΔFPR_2	20.29 ± 0.396	11.31 ± 0.354	0.00 ± 0.362	14.48 ± 0.262	1.88 ± 0.491	**23.65 ± 0.309**
	ΔFPR_3	20.79 ± 0.514	11.08 ± 0.642	0.00 ± 0.466	13.44 ± 0.334	1.21 ± 0.674	**23.15 ± 0.436**
	ΔEER	**12.49 ± 0.179**	4.88 ± 0.130	1.09 ± 0.173	2.58 ± 0.031	0.00 ± 0.039	7.66 ± 0.142
D05	ΔFPR_1	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000
	ΔFPR_2	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000
	ΔFPR_3	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000
	ΔEER	0.00 ± 0.082	0.58 ± 0.093	3.68 ± 0.152	1.98 ± 0.059	2.38 ± 0.119	**6.75 ± 0.131**
D06	ΔFPR_1	**11.36 ± 0.366**	8.00 ± 0.412	4.12 ± 0.454	2.60 ± 0.355	0.00 ± 0.502	2.74 ± 0.504
	ΔFPR_2	**14.21 ± 0.405**	10.00 ± 0.436	4.89 ± 0.408	3.06 ± 0.310	0.00 ± 0.433	0.82 ± 0.466
	ΔFPR_3	**7.98 ± 0.120**	6.21 ± 0.263	3.55 ± 0.214	2.47 ± 0.095	0.00 ± 0.125	3.63 ± 0.223
	ΔEER	**9.86 ± 0.291**	7.28 ± 0.334	4.94 ± 0.293	4.08 ± 0.273	2.45 ± 0.319	0.00 ± 0.385

speakers than from female speakers. D03, *i.e.*, Spec-ResNet also has higher misclassification rate for male speakers in 20s, 30s and 60s than female speakers. However, this bias is not as high as other detectors. One interesting point to note here is that if for a detector the FPR is 100% for both bona fide speech from male and female speaker, then all ΔFPR metrics will be zero. This does not imply that the detector is not biased, but in such cases ΔEER will be able to show bias. This behaviour is observed for detector D05 *i.e.*, LFCC-GMMs, stressing on the fact that simple GMMs based synthetic speech detectors may not be suitable for deployment for in-the-wild synthetic speech detection scenarios (also observed in [15]). Overall, it can be observed from Tab. 3 that majority of the approaches misclassify bona fide speech from male speakers more than they do for female speakers. Coincidently, similar bias for male gender was also found in deepfake face detectors [34].

4.3. Experiment 3: Studying Bias on Age

In this experiment, we determine if a detector misclassifies bona fide speech from one particular age group more than other age groups. We fixed the accent to US English and studied age bias on two genders male and female. For each gender, we have six sets having bona fide speech from speakers in teens, 20s, 30s, 40s, 50s, and 60s.

Tab. 4 and Tab. 5 show results from our age bias study on male and female gender, respectively. We observe that majority of the detectors have more misclassification for speakers in extreme age groups *i.e.*, speakers who are either teens or in 60s. In contrast to this, detector D04 misclassifies bona fide speech from speakers in their 30s more than speakers in other age groups. This behaviour is only observed for male speakers (Tab. 4) and not for female speakers (Tab. 5). For both male and female groups, detector D03 [44] has least bias among different age groups. Detector D05, *i.e.*, LFCC-GMMs again has 100% FPR on all accents, resulting in ΔFPR of zero. However, ΔEER shows that this method misclassifies more bona fide speech for speakers in 60s and 30s age groups. This is true for both male and female speakers. Overall, majority of detectors unfairly misclassify bona fide speech from speakers in teens and 60s age group. Depending on detector and bias measure scenario, the bias varies from minimum 0.3 percentage points (D01 in Tab. 5) to approximately 23 percentage points (D04 in Tab. 7).

4.4. Experiment 4: Studying Bias on Accent

In this experiment, we determine if a detector unfairly targets English bona fide speech with one particular accent.

We fixed the age group to speakers in their 20s, and performed an independent accent bias study on male and female genders. For each gender, we performed an accent bias study on five different accents: Canadian (CN), United States (US), British (UK), Australian (AU), and South Asian (SA).

Tab. 6 and Tab. 7 show results from our accent bias study on male and female gender, respectively. The results show that the detectors are most accurate for US accent English. This may be because most of the open source datasets such as ASVspoof2019 [21] and LibriSpeech [63] on which these detectors are fine-tuned or pretrained have majority speech samples with US accent. Almost all detectors have higher FPR for bona fide speech from South Asian (SA) and Australian (AU) English speakers. Also notice the **bold** values in Tab. 6 and Tab. 7. Detector D02 (Wav2Vec2-AASIST [16]) misclassifies bona fide speech from South Asian female speakers approximately 52 percentage points higher than Canadian female speakers. This observation is also true for bona fide speech from South Asian male speakers with approximately 33 percentage points higher misclassification than Australian male speakers. Apart from high FPRs for

Table 6. Accent Bias Study for Male Speakers. Values are in %.

DN	Metric	$D_{CN-20s-M}$	$D_{US-20s-M}$	$D_{UK-20s-M}$	$D_{AU-20s-M}$	$D_{SA-20s-M}$
D01	ΔFPR_1	0.71 ± 0.070	0.40 ± 0.180	0.29 ± 0.177	0.00 ± 0.020	**0.74 ± 0.106**
	ΔFPR_2	0.05 ± 0.012	0.00 ± 0.015	0.02 ± 0.021	**0.06 ± 0.010**	0.02 ± 0.015
	ΔFPR_3	6.16 ± 0.379	4.76 ± 0.699	0.00 ± 0.497	2.07 ± 0.353	**9.63 ± 0.455**
	ΔEER	5.21 ± 0.227	3.62 ± 0.335	0.00 ± 0.273	5.40 ± 0.195	**9.17 ± 0.343**
D02	ΔFPR_1	5.65 ± 0.214	5.33 ± 0.451	5.09 ± 0.233	0.00 ± 0.045	**33.31 ± 0.456**
	ΔFPR_2	2.90 ± 0.209	2.55 ± 0.334	2.78 ± 0.151	0.00 ± 0.038	**7.31 ± 0.120**
	ΔFPR_3	0.95 ± 0.046	1.74 ± 0.236	1.79 ± 0.224	0.00 ± 0.010	**5.86 ± 0.108**
	ΔEER	1.06 ± 0.026	1.23 ± 0.157	1.50 ± 0.174	0.00 ± 0.018	**4.80 ± 0.112**
D03	ΔFPR_1	0.13 ± 0.076	0.00 ± 0.107	0.11 ± 0.088	**0.14 ± 0.075**	0.02 ± 0.095
	ΔFPR_2	0.13 ± 0.076	0.01 ± 0.097	0.16 ± 0.092	**0.18 ± 0.075**	0.00 ± 0.106
	ΔFPR_3	**0.06 ± 0.026**	0.00 ± 0.036	0.05 ± 0.041	0.04 ± 0.026	0.00 ± 0.055
	ΔEER	**1.78 ± 0.097**	0.96 ± 0.111	1.26 ± 0.110	1.66 ± 0.091	0.00 ± 0.128
D04	ΔFPR_1	0.00 ± 0.108	3.34 ± 0.792	**11.51 ± 0.214**	9.98 ± 0.087	6.09 ± 0.289
	ΔFPR_2	0.00 ± 0.256	2.29 ± 0.315	**11.14 ± 0.252**	9.71 ± 0.185	5.50 ± 0.439
	ΔFPR_3	0.00 ± 0.355	3.32 ± 0.627	**11.74 ± 0.338**	10.29 ± 0.257	5.68 ± 0.310
	ΔEER	0.00 ± 0.258	0.39 ± 0.336	1.40 ± 0.300	**2.98 ± 0.183**	2.49 ± 0.248
D05	ΔFPR_1	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000
	ΔFPR_2	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000
	ΔFPR_3	0.00 ± 0.007	0.00 ± 0.007	0.00 ± 0.007	0.00 ± 0.007	0.00 ± 0.010
	ΔEER	5.05 ± 0.137	5.21 ± 0.256	4.27 ± 0.180	**6.48 ± 0.125**	0.00 ± 0.176
D06	ΔFPR_1	2.26 ± 0.189	0.36 ± 0.333	5.21 ± 0.289	0.00 ± 0.056	**5.30 ± 0.215**
	ΔFPR_2	2.74 ± 0.275	0.91 ± 0.487	**6.38 ± 0.299**	0.00 ± 0.072	6.23 ± 0.291
	ΔFPR_3	2.09 ± 0.535	0.00 ± 0.731	**4.26 ± 0.559**	0.48 ± 0.517	4.17 ± 0.553
	ΔEER	2.07 ± 0.118	0.69 ± 0.119	3.15 ± 0.107	0.00 ± 0.032	**3.09 ± 0.155**

Table 7. Accent Bias Study for Female Speakers. Values are in %.

DN	Metric	$D_{CN-20s-M}$	$D_{US-20s-M}$	$D_{UK-20s-M}$	$D_{AU-20s-M}$	$D_{SA-20s-M}$
D01	ΔFPR_1	0.00 ± 0.048	1.68 ± 0.160	2.79 ± 0.071	**4.16 ± 0.060**	3.97 ± 0.095
	ΔFPR_2	0.00 ± 0.013	0.33 ± 0.038	0.65 ± 0.009	**0.73 ± 0.013**	0.64 ± 0.056
	ΔFPR_3	2.47 ± 0.228	3.97 ± 0.375	0.00 ± 0.304	**14.03 ± 0.240**	13.60 ± 0.358
	ΔEER	1.95 ± 0.068	2.07 ± 0.290	0.00 ± 0.038	**17.61 ± 0.138**	6.02 ± 0.183
D02	ΔFPR_1	0.00 ± 0.070	7.78 ± 0.422	12.39 ± 0.075	40.00 ± 0.207	**52.26 ± 0.595**
	ΔFPR_2	0.00 ± 0.105	1.85 ± 0.184	1.99 ± 0.104	7.51 ± 0.101	**8.99 ± 0.150**
	ΔFPR_3	0.00 ± 0.016	1.40 ± 0.102	3.77 ± 0.065	2.07 ± 0.097	**26.62 ± 0.618**
	ΔEER	0.00 ± 0.000	1.32 ± 0.203	3.21 ± 0.054	2.39 ± 0.061	**16.30 ± 0.389**
D03	ΔFPR_1	0.88 ± 0.053	0.78 ± 0.072	0.97 ± 0.053	**1.06 ± 0.056**	0.00 ± 0.074
	ΔFPR_2	0.95 ± 0.104	0.92 ± 0.145	1.03 ± 0.103	**1.27 ± 0.104**	0.00 ± 0.145
	ΔFPR_3	0.50 ± 0.116	0.46 ± 0.126	0.48 ± 0.116	0.60 ± 0.116	0.00 ± 0.164
	ΔEER	2.35 ± 0.299	2.87 ± 0.324	2.19 ± 0.302	**6.32 ± 0.301**	0.00 ± 0.421
D04	ΔFPR_1	0.00 ± 0.117	9.87 ± 0.473	20.24 ± 0.226	36.94 ± 0.143	25.20 ± 0.560
	ΔFPR_2	0.00 ± 0.080	10.21 ± 0.970	20.52 ± 0.189	**37.07 ± 0.148**	25.80 ± 0.468
	ΔFPR_3	0.00 ± 0.088	9.58 ± 0.467	19.62 ± 0.084	**36.78 ± 0.273**	24.71 ± 0.458
	ΔEER	0.00 ± 0.036	2.71 ± 0.212	7.61 ± 0.079	6.93 ± 0.048	**8.34 ± 0.334**
D05	ΔFPR_1	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000
	ΔFPR_2	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000
	ΔFPR_3	0.02 ± 0.000	0.02 ± 0.000	0.00 ± 0.000	0.02 ± 0.000	0.02 ± 0.000
	ΔEER	10.98 ± 0.088	9.77 ± 0.187	**11.46 ± 0.118**	11.00 ± 0.103	0.00 ± 0.121
D06	ΔFPR_1	5.86 ± 0.149	5.22 ± 0.498	0.00 ± 0.210	10.71 ± 0.165	**8.68 ± 0.240**
	ΔFPR_2	6.23 ± 0.100	5.64 ± 0.541	0.00 ± 0.089	**11.21 ± 0.184**	9.51 ± 0.642
	ΔFPR_3	5.80 ± 0.117	4.33 ± 0.247	0.00 ± 0.151	**8.50 ± 0.138**	7.00 ± 0.326
	ΔEER	1.59 ± 0.074	1.62 ± 0.199	0.00 ± 0.089	**4.70 ± 0.152**	2.69 ± 0.276

bona fide speech from Australian and South Asian speakers, detector $D04$, (PS3DT) [23] also has around 11 percentage points higher misclassification for bona fide speech with British accents than Canadian accent Tab. 6. Overall, the majority of detectors unfairly target bona fide speech from speakers with South Asian and Australian accents.

4.5. Experiment 4: Bias on Stuttering Speech

In this experiment, we examine the performance of existing synthetic speech detectors on bona fide speech from impaired speakers, particularly with stuttering.

From Tab. 2, the mean EER of all the detectors on D_{eval} set which contains bona fide speech from fluent speakers and synthetic speech from 11 unknown and 2 known speech generators is 5.29%. However, results from Tab. 8 show that the mean EER from each detector increases and become 36.31% when the bona fide speech in D_{eval} set is replaced with bona fide speech with stuttering from Sep-28k [36, 37] dataset. Note: The EER rate is same as FPR and FNR as EER is the rate that makes FPR and FNR equal. Therefore, we can conclude that in the scenario where ground labels are given and EER is calculated, the mean $FPR@EER$ Threshold would also increase from 5.29% to 36.31% for bona fide speech from impaired speakers (also shown in Fig. 1).

Table 8. Stuttering Speech Bias Study. Values are in %.

DN	FPR_1	FPR_2	FPR_3	EER
D01	88.10	96.54	66.13	34.10
D02	69.96	96.77	24.99	18.16
D03	96.40	95.58	97.75	47.19
D04	52.02	53.09	50.68	22.40
D05	97.32	98.53	94.06	49.05
D06	87.44	83.41	92.18	46.94
Mean	81.87	87.32	70.97	36.31

In the scenarios where ground truth labels are not provided, and FPR is measured using pre-determined thresholds obtained from an independent dataset, we can observe that mean FPRs are always greater than 70%. Therefore all detectors are biased and often misclassify bona fide speech with stuttering as synthetic more than fluent bona fide speech.

5. Conclusions

In this work, we examined bias in different methods for synthetic speech detection. We processed a corpus of 1.7 million bona fide speech to create 28 different demographic sets. We evaluated age, gender and accent bias in 6 synthetic speech detectors. Results indicate that synthetic speech detectors are biased. False positive rates are higher for bona fide speech from male gender, speakers in age groups teens and 60s, and speakers with South Asian and Australian English accents in comparison to other demographic groups. We also found that synthetic speech detectors are unfair to speech-impaired speakers. Future work will focus on developing unbiased synthetic speech detectors.

Acknowledgements This material is based on research sponsored by the Defense Advanced Research Projects Agency (DARPA) and the Air Force Research Laboratory (AFRL) under agreement number FA8750-20-2-1004. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA, AFRL or the U.S. Government. Address all correspondence to Edward J. Delp, ace@purdue.edu.

References

[1] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection," *Proceedings of the ISCA Interspeech*, pp. 1008–1012, September 2019, Graz, Austria. 1, 2, 3, 5, 6

[2] K. Bhagtani, A. K. S. Yadav, E. R. Bartusiak, Z. Xiang, R. Shao, S. Baireddy, and E. J. Delp, "An Overview of Recent Work in Multimedia Forensics," *Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval*, pp. 324–329, August 2022, Virtual. 1

[3] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, December 2020. 1

[4] P. Dhariwal and A. Q. Nichol, "Diffusion Models Beat GANs on Image Synthesis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8780–8794, December 2021, Virtual. 1

[5] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10 684–10 695, 2022. 1

[6] "Speech Synthesis, ElevenLabs," December 2023. [Online]. Available: https://elevenlabs.io/ 1

[7] H. Kim, S. Kim, J. Yeom, and S. Yoon, "UnitSpeech: Speaker-adaptive Speech Synthesis with Untranscribed Data," *Proceedings of the ISCA Interspeech*, pp. 3038–3042, August 2023, Dublin, Ireland. 1

[8] Coqui, "XTTS," September 2023. [Online]. Available: https://docs.coqui.ai/en/latest/models/xtts.html 1

[9] R. Huang, Z. Zhao, H. Liu, J. Liu, C. Cui, and Y. Ren, "ProDiff: Progressive Fast Diffusion Model for High-Quality Text-to-Speech," *Proceedings of the ACM International Conference on Multimedia*, pp. 2595–2605, October 2022, Lisbon, Portugal. 1

[10] E. Flitter and S. Cowley, "Voice Deepfakes Are Coming for Your Bank Balance." *The New York Times*, August 2023. [Online]. Available: https://www.nytimes.com/2023/08/30/business/voice-deepfakes-bank-scams.html 1

[11] B. Nguyen, "A couple in Canada were reportedly scammed out of $21,000 after getting a call from an AI-generated voice pretending to be their son." *The New York Times*, March 2023. [Online]. Available: https://www.businessinsider.com/couple-canada-reportedly-lost-21000-in-ai-generated-voice-scam-2023-3 1

[12] B. MOLLMAN, "Scammers are using voice-cloning A.I. tools to sound like victims' relatives in desperate need of financial help. It's working." *The New York Times*, March 2023. [Online]. Available: https://fortune.com/2023/03/05/scammers-ai-voice-cloning-tricking-victims-sound-like-relatives-needing-money/ 1

[13] P. Verma, "They thought loved ones were calling for help. It was an AI scam." *The Washington Post*, March 2023. [Online]. Available: https://www.washingtonpost.com/technology/2023/03/05/ai-voice-scam/ 1

[14] G. Hua, A. B. J. Teoh, and H. Zhang, "Towards End-to-End Synthetic Speech Detection," *IEEE Signal Processing Letters*, vol. 28, pp. 1265–1269, June 2021. 1, 2, 3, 4, 6

[15] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch *et al.*, "Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild," *arXiv preprint*, 2022. 1, 2, 4, 7

[16] H. Tak, M. Todisco, X. Wang, J. weon Jung, J. Yamagishi, and N. Evans, "Automatic Speaker Verification Spoofing and Deepfake Detection Using Wav2vec 2.0 and Data Augmentation," *Proceedings of the Speaker and Language Recognition Workshop, Odyssey*, pp. 112–119, July 2022, Beijing, China. 1, 2, 3, 4, 6, 7

[17] K. Li, X.-M. Zeng, J.-T. Zhang, and Y. Song, "Convolutional recurrent neural network and multitask learning for manipulation region location," *Proceedings of IJCAI Workshop on Deepfake Audio Detection and Analysis*, pp. 18–22, August 2023, Macao. 1, 2

[18] Z. Zhang, X. Yi, and X. Zhao, "Fake Speech Detection Using Residual Network with Transformer Encoder," *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, p. 13–22, June 2021, virtual Event, Belgium. 1, 2

[19] C. Sun, S. Jia, S. Hou, and S. Lyu, "Ai-synthesized voice detection using neural vocoder artifacts," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 904–912, June 2023, Vancouver, Canada. 1, 2

[20] Z. Xiang, A. K. S. Yadav, S. Tubaro, P. Bestagini, and E. J. Delp, "Extracting efficient spectrograms from mp3 compressed speech signals for synthetic speech detection," *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, p. 163–168, 2023, Chicago, IL, USA. 1

[21] J. Yamagishi, M. Todisco, M. Sahidullah, H. Delgado, X. Wang, N. Evans, T. Kinnunen, K. Lee, V. Vestman, and A. Nautsch, "ASVspoof 2019: The 3rd Automatic Speaker Verification Spoofing and Countermeasures Challenge database," *University of Edinburgh. The Centre for Speech Technology Research*, March 2019. [Online]. Available: https://www.asvspoof.org/index2019.html 1, 2, 3, 4, 7

[22] N. M. Müller, P. Czempin, F. Dieckmann, A. Froghyar, and K. Böttinger, "Does audio deepfake detection generalize?" *Proceedings of the ISCA Interspeech*, September 2022, Incheon, Korea. 1, 2

[23] A. K. Singh Yadav, Z. Xiang, K. Bhagtani, P. Bestagini, S. Tubaro, and E. J. Delp, "PS3DT: Synthetic Speech Detection Using Patched Spectrogram Transformer," *Proceedings of the IEEE International Conference on Machine Learning and Applications*, pp. 496–503, 2023, Florida, USA. 1, 2, 3, 6, 8

[24] W. Ge, J. Patino, M. Todisco, and N. Evans, "Explaining Deep Learning Models for Spoofing and Deepfake Detection with Shapley Additive Explanations," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6387–6391, May 2022, Singapore. 1, 2

[25] S.-Y. Lim, D.-K. Chae, and S.-C. Lee, "Detecting Deepfake Voice Using Explainable Deep Learning Techniques," *Applied Sciences*, vol. 12, no. 8, 2022. 1

[26] H. Tak, J. Patino, A. Nautsch, N. Evans, and M. Todisco, "An explainability study of the constant Q cepstral coefficient spoofing countermeasure for automatic speaker verification ," *Proceedings of the Speaker and Language Recognition Workshop*, pp. 333–340, November 2020, Tokyo, Japan. 1

[27] D. Salvi, P. Bestagini, and S. Tubaro, "Towards frequency band explainability in synthetic speech detection," in *European Signal Processing Conference (EUSIPCO)*. IEEE, 2023. 1

[28] A. K. Singh Yadav, K. Bhagtani, Z. Xiang, P. Bestagini, S. Tubaro, and E. J. Delp, "DSVAE: Disentangled Representation Learning for Synthetic Speech Detection," *Proceedings of the IEEE International Conference on Machine Learning and Applications*, pp. 472–479, 2023, Florida, USA. 1, 2

[29] A. K. Singh Yadav, Z. Xiang, E. R. Bartusiak, P. Bestagini, S. Tubaro, and E. J. Delp, "ASSD: Synthetic Speech Detection in the AAC Compressed Domain," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 1–5, June 2023, Rhodes Island, Greece. 1, 2, 3

[30] S. Feng, O. Kudina, B. M. Halpern, and O. Scharenborg, "Quantifying bias in automatic speech recognition," *arXiv preprint arXiv:2103.15122*, March 2021. 1, 2, 5

[31] W. T. Hutiri and A. Y. Ding, "Bias in automated speaker recognition," *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, p. 230–247, June 2022, Seoul, Republic of Korea. 1, 2, 5

[32] M. K. Ngueajio and G. Washington, "Hey ASR System! Why Aren't You More Inclusive?" *Proceedings of International Conference on Human-Computer Interaction*, November 2022. 1, 2

[33] Y. Xu, P. Terhörst, K. Raja, and M. Pedersen, "A comprehensive analysis of ai biases in deepfake detection with massively annotated databases," *arXiv preprint arXiv:2208.05845*, September 2023. 1, 2, 3

[34] M. Pu, M. Y. Kuan, N. T. Lim, C. Y. Chong, and M. K. Lim, "Fairness evaluation in deepfake detection models using metamorphic testing," *Proceedings of the ACM International Workshop on Metamorphic Testing*, p. 7–14, January 2023, Pittsburgh, Pennsylvania. 1, 2, 3, 7

[35] M. Organization, "Common Voice Corpus 16.1." January 2024. [Online]. Available: https://commonvoice.mozilla.org/en/datasets 2, 4

[36] C. Lea, V. Mitra, A. Joshi, S. Kajarekar, and J. Bigham, "Sep-28k: A dataset for stuttering event detection from podcasts with people who stutter," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6798–6802, June 2021, Toronto, Canada. 2, 5, 8

[37] S. P. Bayerl, D. Wagner, T. Bocklet, and K. Riedhammer, "The Influence of Dataset-Partitioning on Dysfluency Detection Systems," *Text, Speech, and Dialogue*, vol. 13502, 2022. 2, 5, 8

[38] M. Sahidullah and G. Saha, "Design, Analysis, and Experimental Evaluation of Block Based Transformation in MFCC Computation for Speaker Recognition," *Speech Communication*, vol. 54, pp. 543–565, May 2012. 2, 3

[39] M. Todisco, H. Delgado, and N. Evans, "Constant Q Cepstral Coefficients: A Spoofing Countermeasure for Automatic Speaker Verification," *Computer Speech & Language*, vol. 45, pp. 516–535, September 2017. 2

[40] S. Borzì, O. Giudice, F. Stanco, and D. Allegra, "Is synthetic voice detection research going into the right direction?" *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 71–80, June 2022, New Orleans, USA. 2

[41] X. Li, N. Li, C. Weng, X. Liu, D. Su, D. Yu, and H. Meng, "Replay and Synthetic Speech Detection with Res2Net Architecture," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6354–6358, June 2021, Toronto, Canada. 2

[42] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," *Proceedings of the ISCA Interspeech*, pp. 2087–2091, September 2015, Dresden, Germany. 2, 3

[43] F. Akdeniz and Y. Becerikli, "Detection of Copy-Move Forgery in Audio Signal with Mel Frequency and Delta-Mel Frequency Kepstrum Coefficients," *Proceedings of the Innovations in Intelligent Systems and Applications Conference*, pp. 1–6, October 2021, Elazig, Turkey. 2

[44] M. Alzantot, Z. Wang, and M. B. Srivastava, "Deep Residual Neural Networks for Audio Spoofing Detection," *Proceedings of the ISCA Interspeech*, pp. 1078–1082, September 2019, Graz, Austria. 2, 3, 6, 7

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, June 2016, Las Vegas, NV. 2, 3

[46] H. Zeinali, T. Stafylakis, G. Athanasopoulou, J. Rohdin, I. Gkinis, L. Burget, and J. Černockỳ, "Detecting Spoofing Attacks Using VGG and SincNet: BUT-Omilia Submission to ASVspoof 2019 Challenge," *Proceedings of the ISCA Interspeech*, pp. 1073–1077, September 2019, Graz, Austria. 2

[47] E. R. Bartusiak, K. Bhagtani, A. K. S. Yadav, and E. J. Delp, "Transformer Ensemble for Synthesized Speech Detection," *Proceedings of the Asilomar Conference on Signals, Systems, and Computers*, October 2023, Pacific Grove, California, USA. 2, 3

[48] S. S. Stevens, J. Volkmann, and E. B. Newman, "A Scale for the Measurement of the Psychological Magnitude Pitch," *Journal of the Acoustical Society of America*, vol. 8, pp. 185–190, June 1937. 2, 3

[49] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013. 2

[50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Łukasz Kaiser, and I. Polosukhin, "Attention is All You Need," *Proceedings of the Neural Information Processing Systems*, December 2017, Long Beach, CA. 2, 4

[51] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, "Efficient Training of Audio Transformers with Patchout," *Proceedings of the ISCA Interspeech*, pp. 2753–2757, September 2022, Incheon, Korea. 2

[52] Y. Gong, C.-I. Lai, Y.-A. Chung, and J. Glass, "SSAST: Self-Supervised Audio Spectrogram Transformer," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, pp. 10 699–10 709, October 2022, Virtual. 2

[53] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," *Proceedings of the ISCA Interspeech*, pp. 571–575, August 2021, Brno, Czech Republic. 2

[54] A. K. S. Yadav, E. Bartusiak, K. Bhagtani, and E. J. Delp, "Synthetic Speech Attribution using Self Supervised Audio Spectrogram Transformer," *Proceedings of the IS&T Media Watermarking, Security, and Forensics Conference, Electronic Imaging Symposium*, January 2023, san Francisco, CA. 2

[55] K. Bhagtani, E. R. Bartusiak, A. K. S. Yadav, P. Bestagini, and E. J. Delp, "Synthesized speech attribution using the patchout spectrogram attribution transformer," *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, p. 157–162, June 2023, Chicago, IL, USA. 2

[56] A. K. Singh Yadav, K. Bhagtani, S. Baireddy, P. Bestagini, S. Tubaro, and E. J. Delp, "Mdrt: Multi-domain synthetic speech localization," *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 11 171–11 175, 2024, Seoul, South Korea. 2

[57] K. Bhagtani, A. K. S. Yadav, Z. Xiang, P. Bestagini, and E. J. Delp, "FGSSAT : Unsupervised Fine-Grain Attribution of Unknown Speech Synthesizers Using Transformer Networks," *Proceedings of the IEEE Asilomar Conference on Signals, Systems, and Computers*, pp. 1135–1140, 2023, Pacific Grove, CA. 2

[58] L. Zhang, X. Wang, E. Cooper, and J. Yamagishi, "Multi-task Learning in Utterance-level and Segmental-level Spoof Detection," *Proceedings of the Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, pp. 9–15, September 2021, Online. 2

[59] Y. Zhu, Y. Chen, Z. Zhao, X. Liu, and J. Guo, "Local self-attention based hybrid multiple instance learning for partial spoof speech detection," *ACM Transactions on Intelligent Systems and Technology*, August 2023. 2

[60] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," October 2020. 2, 4

[61] L. Zhang, X. Wang, E. Cooper, N. Evans, and J. Yamagishi, "The partialspoof database and countermeasures for the detection of short fake speech segments embedded in an utterance," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 813–825, 2023. 2

[62] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An Ontology and Human-labeled Dataset for Audio Events," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2017, New Orleans, LA. 2, 3

[63] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: an ASR Corpus Based on Public Domain Audio Books," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2015, Queensland, Australia. 2, 7

[64] K. Kärkkäinen and J. Joo, "Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation," *Proceedings of IEEE Winter Conference on Applications of Computer Vision*, pp. 1547–1557, January 2021, Hawaii, USA. 2

[65] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–35, July 2021. 2

[66] C. Hazirbas, J. Bitton, B. Dolhansky, J. Pan, A. Gordo, and C. C. Ferrer, "Towards measuring fairness in ai: The casual conversations dataset," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 4, no. 3, pp. 324–332, 2022. 2

[67] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, "Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward," *Applied Intelligence*, vol. 53, no. 4, pp. 3974–4026, June 2023. 2, 3

[68] R. Ramachandra, K. Raja, and C. Busch, "Algorithmic fairness in face morphing attack detection," *Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, pp. 410–418, January 2022, Hawaii, USA. 2

[69] M. Fang, W. Yang, A. Kuijper, V. Struc, and N. Damer, "Fairness in face presentation attack detection," *Pattern Recognition*, vol. 147, p. 110002, October 2023. 2

[70] L. Trinh and Y. Liu, "An examination of fairness of ai models for deepfake detection," *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 567–574, August 2021, Montreal, Canada. 2

[71] A. V. Nadimpalli and A. Rattani, "Gbdf: Gender balanced deepfake dataset towards fair deepfake detection," *arXiv preprint arXiv:2207.10246*, July 2022. 3

[72] Y. Ju, S. Hu, S. Jia, G. H. Chen, and S. Lyu, "Improving fairness in deepfake detection," *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pp. 4655–4665, January 2024, Hawaii, USA. 3

[73] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Masked Spectrogram Modeling using Masked Autoencoders for Learning General-purpose Audio Representation," *Proceedings of Machine Learning Research*, vol. 166, pp. 1–24, Dec 2022. 3

[74] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6367–6371, May 2022, Singapore. 3, 4

[75] W. Ge, J. Patino, M. Todisco, and N. Evans, "Raw Differentiable Architecture Search for Speech Deepfake and Spoofing Detection," *Proceedings of the ISCA Automatic Speaker Verification and Spoofing Countermeasures Challenge*, pp. 22–28, 2021. 3