Supplementary Material for "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

1. Detectors Training Details

Below we provide additional training details for the detectors used in our study.

1.1. LFCC-GMMs

While this method in [1] is implemented in MATLAB, we used scikit learn to fit the Gaussian Mixture Models (GMMs). Tab. 1 presents our ablation study. The M01 version reports the performance of the method trained in MATLAB using the parameters provided for the baseline in ASVspoof2019 Challenge [2]. M01 processes 20ms windows with a hop size of 10ms and frequency components from 30Hz to 8kHz to obtain Linear Frequency Cepstral Coefficients (LFCCs) features. The M02 version reports the performance of the method using same parameters but trained in Python using scikit learn. The M03 version is implemented in scikit learn and processes 30ms windows with a hop size of 15ms and frequency components upto 4kHz to obtain LFCCs features. Tab. 1 shows the Equal Error Rate (EER) in % of all three versions on the D_{eval} set of the ASVspoof2019 dataset. Two GMMs are trained separately, each one for the bona fide and synthetic classes using Expectation Maximization (EM) algorithm. Each GMM has 512 Gaussian mixture components. The covariance matrix of the Gaussian distributions was constrained to be a diagonal matrix. For all three versions, we obtained 20 LFCCs, their deltas and also delta-delta coefficients as described in [2]. Therefore, we obtain 60-dimensional features for each small window of the speech signal (20ms in case of M01 and M02, and 30ms in case of M03). Since window size in M02 is smaller, we obtained a higher number of windows in case of M02 as compared to that of M03. We observed that M02 was relatively more computationally intensive than M03, and was slow to train but both versions had comparable performance on the evaluation set of the ASVspoof2019 dataset as shown in Tab. 1. In our study, we required to evaluate this method on 0.9 million speech signals, which is why we selected M03 for our study to examine bias in LFCC-GMMs [1, 3] synthetic speech detector.

Table 1. Ablation on LFCC-GMMs showing EER in % on D_{eval}.

Version	Window Size	Hop Size	Package	D_{eval}
M01	20ms	10ms	MATLAB	8.09
M02	20ms	10ms	scikit-learn	3.67
M03	30ms	15ms	scikit-learn	3.46

Table 2. Ablation on MFCC-ResNet showing EER in % on D_{dev} and D_{eval}.

Method Version	D_{dev}	D_{eval}
Presented in [4]	3.34	9.33
Using provided model weights	>40.00	>40.00
Our retrained version	6.52	11.58

1.2. MFCC-ResNet

The features used by this detector consist of Mel-Frequency Cepstrum Coefficients (MFCC), and its first and second derivatives. The MFCC are obtained using the Short Time Fourier Transform (STFT) of the speech signal, mel-spectrum filters and Discrete Cosine Transform (DCT) [4]. We select the first 24 coefficients. Including, its first and second derivatives, in total, it leads to a feature of dimension 72. We use these parameters because they have the best performance in [4]. These hand-crafted features are processed by a ResNet [5] provided in [4] for synthetic speech detection.

The weights provided with [4] were for a model which was trained on MFCC features obtained using the librosa.feature package. However, later (after the source code and model weights were released), the package was updated which changed the number of windows obtained for each speech signal. Hence, corresponding changes were needed to be

Table 3. Ablation on Spec-ResNet showing EER in % on D_{dev} and D_{eval}.

Method Version	D_{dev}	D_{eval}
Presented in [4]	0.11	9.68
Using provided model weights	43.05	42.01
Our retrained version	0.71	10.10

made in the ResNet architecture provided by the authors in the source code. After making these number of windows changes, in our first experiment, we evaluated the performance of the model weights provided by the authors. There was a significant drop in performance as compared to the results presented in [4]. Using the model weights provided by the authors, we observed an EER higher than 40% on both development D_{dev} and evaluation D_{eval} sets of the ASVspoof2019 Dataset. Hence, we retrained the method on the ASVspoof2019 dataset and obtained our own model weights.

We trained this method for 200 epochs using a learning rate of 5×10^{-5} and a batch size of 32. We selected the model which performed the best on the validation set of ASVspoof2019. In Tab. 2, the EER in % on D_{dev} and D_{eval} sets of the ASVspoof2019 dataset obtained by using our own model weights are shown. This led to better performance than directly using the model weights provided by the authors, hence we used the re-trained version for our bias study.

1.3. Spec-ResNet

For similar reasons as mentioned for MFCC-ResNet, we re-trained this method. This method obtains a 2048-point STFT from the speech signal using a window size of 2048 and hop length of 512. Next, the squares of absolute value of the STFT are obtained and converted into decibles (dB) scale, which is a logarithmic scale. We trained this method for 200 epochs using using a learning rate of 5×10^{-5} and a batch size of 32. The model with the best performance on the validation set was selected. The EER in % on D_{dev} and D_{eval} sets of ASVspoof2019 obtained using model weights provided in [4] and our re-trained model are shown in Tab. 3. We observed that our re-training helped to reduce EER significantly. Hence, we use our retrained model weights for the bias study.

1.4. PS3DT

To obtain mel-spectrogam, we used a Hanning window of size 25 ms with a shift of 10 ms. As mentioned in [6], we used 80 frequency bins and fixed input speech signal to 5.12 seconds, resulting in a mel-spectrogram of size 80×512. The network was trained for 50 epochs with a batch size of 256 and AdamW optimizer [7]. The initial learning rate was set to 10^{-5} and a weight decay of 10^{-4} was used. We selected the model weights which provide best accuracy on the validation set. We obtained performance same as mentioned in [6].

1.5. TSSDNet

We do not perform any re-training for this method. We performed evaluation using the ResNet style Time-Domain Synthetic Speech Detection Network (TSSDNet) model weights provided in [8] as the results obtained with it were same as reported in [8].

1.6. Wav2Vec2-AASIST

We do not perform any re-training for this method. We used the weights provided by the authors in [9] for this method as the results obtained with it were same as reported in [9].

2. Dataset Collection and Pre-Processing

In this section, we provide details about how we collected dataset for our bias study and how we processed it. For our age, gender and bias studies, we created 28 evaluation sets. Each evaluation set has bona fide class and synthetic class as shown in Fig. 1. We kept synthetic class same in each set. The sets only differ in terms of bona fide speech signals and particularly bona fide speakers' demographics. For collecting samples for bona fide class we downloaded the Mozilla Common Voice Corpus 16.1 [10]. It has approximately 1.78 million English speech signals. We pre-processed the dataset and obtained approximately 0.9 million speech signals having all the required annotations for our study. To pre-process, we used the annotations provided in the dataset. To handle large number of files and its processing, we run parallel process using GNU parallel [11]. We used all the speech signal with gender, age and accent annotations. We considered only annotations which are validated. Also, since each detector is trained on 16KHz samples. We resampled each speech signal from the Mozilla Common Voice Corpus (mostly 44KHz) to 16KHz.

For gender, there are three categories: male, female and others in the Mozilla Common Voice Corpus. In gender bias study, we fixed the accent to US English as it had most number of samples available. We studied three most frequent age groups in the dataset, namely, 20s, 30s and 60s. For each group, we made two sets: one for male and other for female. We kept same number of samples for each gender in gender bias study. We limited our categories to only male and female as adding other gender samples lead to strong unbalance in the dataset and for fair evaluation, we wanted to keep number of samples same for both genders in a particular age group.

For age, we fixed accent to US English and the age categories in Mozilla Common Voice Corpus are teens, 20s, 30s, 40s, 50s, 60s, 70s, 80s and 90s. We studied gender bias

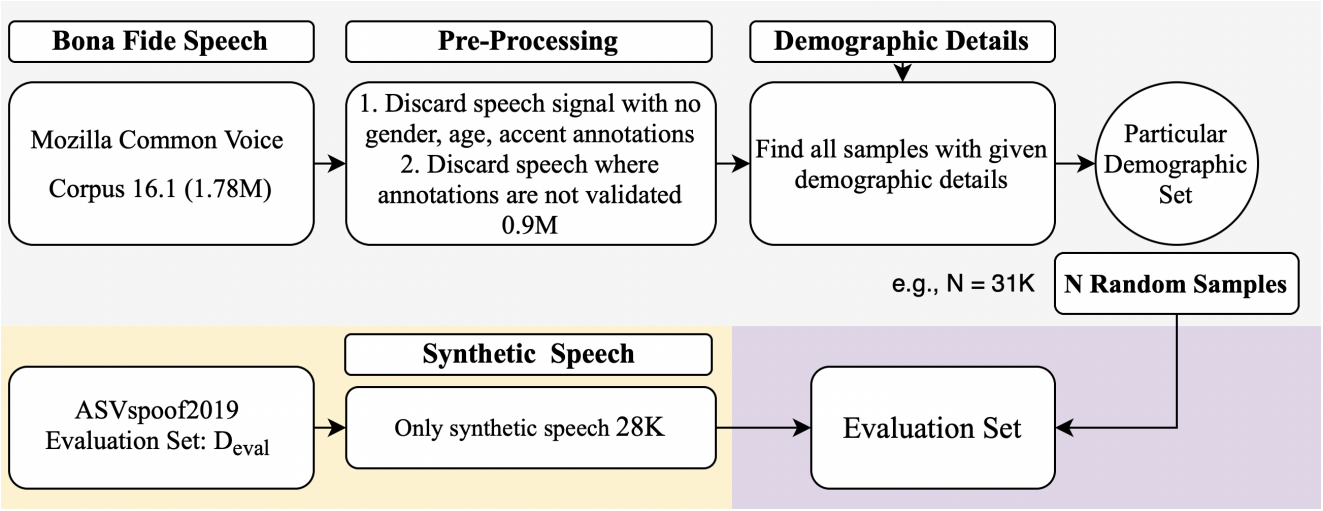

Figure 1. Overview of Dataset Preparation for bias study.

in both male and female gender. For a fixed gender, we kept number of samples same for all age groups. We discarded age groups 70s, 80s, and 90s as they had limited number of samples.

For accent, we fixed age group to 20s as it has the most number of samples. We examined accent bias for male and female gender separately. There are more than 100 different labels for accents. However, many labels have even less than 100 speech samples. In our study, we selected 5 most frequent accents, namely, US English, Canadian English, British English, Australian English, South Asian English. The number of samples are kept same for all the accents. Note, in future work, we plan to provide more fine-grained analysis *e.g.* within bias among accents with US English like Midwestern Accent, California Accent, and so on.

For each evaluation set, we first obtain a bigger set consisting of bona fide speech with a particular demographic as shown in Fig. 1. For example, in our gender study for demographic: speakers in 20s with US Accent, we get a bigger set with 31,500 speech signals for female and 109,000 speech signals for male. From both sets, we randomly sample 'N' (in this case: 'N' = 31K) speech signals as shown in Fig. 1. This brings randomness in our experiment, therefore we report mean metrics and standard deviation obtained from 5 runs of each experiment. For a given demographic, higher standard deviation will show that the results are highly dependent on the content than the demographic itself. We obtained less than 0.05% standard deviation for all metrics. This indicates that the detectors and the results are not dependent on content but demographics in all our experiments. The 28 evaluation datasets used in our age, gender and bias study and script we created to process the dataset can be found at https://gitlab.com/viper-purdue/fairssd. We believe this will help future research in this direction.

3. Absolute Value of FPR and EER

In this section we provide the absolute values of all the metrics. Notice, each experiment is repeated 5 times to get the value. We report both mean and standard deviation (SD). Tab. 4 shows our gender bias study results. Note: D01 is TSSDNet [8], D02 is Wav2Vec2-AASIST (Wav2Vec2) [9], D03 is Spec-ResNet [4], D04 is PS3DT [6], D05 is LFCC-GMMs [1, 2], and D06 is MFCC-ResNet [4]. Similar to the results reported in the paper using difference, we can notice that FPRs are higher for males than for female counterparts for most detectors. Similarly, in Tab. 5 and Tab. 6 we report our age bias study results for male and female genders. Most detectors have higher bias (FPR and EER) for people in age groups teens and 60s. Finally, in Tab. 7 and Tab. 8, we report results for accents. Most detectors have higher FPRs and EERs for speakers with South Asian and Australian English accents.

Notice, while the ΔFPRs and ΔEER reported in the paper do not reveal actual performance, the absolute values reported here reveal absolute performance. We observe that LFFC-GMMs *i.e.*, D05 has 100% FPRs with threshold estimated from an independent dataset *i.e.* D_{eval} set of ASVspoof2019. Hence, this method does not generalize on bona fide speech from unknown speakers and always misclassifies them as synthetic. However, the EER can still show bias. As both classes in each evaluation set have same results on synthetic class (as synthetic class is same in each set). Hence, higher EER for one demographic group indicates method outputs higher probability for bona fide speech from one demographic than that for bona fide speech from other demographic group.

Table 4. Absolute performance of detectors in gender bias study.

Method	Data	Metric	Mean	SD	Data	Metric	Mean	SD	Data	Metric	Mean	SD
	$D_{US-20s-M}$	FPR_1	98.26%	0.086%	$D_{US-30s-M}$	FPR_1	98.59%	0.032%	$D_{US-60s-M}$	FPR_1	98.53%	0.013%
	$D_{US-20s-F}$	FPR_1	96.79%	0.008%	$D_{US-30s-F}$	FPR_1	97.67%	0.013%	$D_{US-60s-F}$	FPR_1	91.97%	0.075%
D01	$D_{US-20s-M}$	FPR_2	99.95%	0.007%	$D_{US-30s-M}$	FPR_2	99.98%	0.011%	$D_{US-60s-M}$	FPR_2	99.91%	0.000%
	$D_{US-20s-F}$	FPR_2	99.59%	0.005%	$D_{US-30s-F}$	FPR_2	99.88%	0.004%	$D_{US-60s-F}$	FPR_2	99.94%	0.006%
	$D_{US-20s-M}$	FPR_3	82.59%	0.157%	$D_{US-30s-M}$	FPR_3	83.26%	0.273%	$D_{US-60s-M}$	FPR_3	91.89%	0.034%
	$D_{US-20s-F}$	FPR_3	81.78%	0.038%	$D_{US-30s-F}$	FPR_3	77.25%	0.071%	$D_{US-60s-F}$	FPR_3	65.77%	0.246%
	$D_{US-20s-M}$	EER	46.57%	0.050%	$D_{US-30s-M}$	EER	43.25%	0.170%	$D_{US-60s-M}$	EER	57.88%	0.020%
	$D_{US-20s-F}$	EER	45.12%	0.018%	$D_{US-30s-F}$	EER	44.46%	0.017%	$D_{US-60s-F}$	EER	43.04%	0.059%
	$D_{US-20s-M}$	FPR_1	27.04%	0.046%	$D_{US-30s-M}$	FPR_1	21.36%	0.310%	$D_{US-60s-M}$	FPR_1	28.71%	0.043%
	$D_{US-20s-F}$	FPR_1	29.13%	0.033%	$D_{US-30s-F}$	FPR_1	23.76%	0.036%	$D_{US-60s-F}$	FPR_1	17.31%	0.204%
D02	$D_{US-20s-M}$	FPR_2	90.70%	0.131%	$D_{US-30s-M}$	FPR_2	83.45%	0.379%	$D_{US-60s-M}$	FPR_2	93.10%	0.032%
	$D_{US-20s-F}$	FPR_2	91.31%	0.018%	$D_{US-30s-F}$	FPR_2	90.19%	0.034%	$D_{US-60s-F}$	FPR_2	92.72%	0.096%
	$D_{US-20s-M}$	FPR_3	2.67%	0.112%	$D_{US-30s-M}$	FPR_3	2.29%	0.090%	$D_{US-60s-M}$	FPR_3	1.47%	0.016%
	$D_{US-20s-F}$	FPR_3	2.35%	0.014%	$D_{US-30s-F}$	FPR_3	1.35%	0.029%	$D_{US-60s-F}$	FPR_3	0.69%	0.009%
	$D_{US-20s-M}$	EER	3.87%	0.038%	$D_{US-30s-M}$	EER	3.66%	0.080%	$D_{US-60s-M}$	EER	3.03%	0.016%
	$D_{US-20s-F}$	EER	3.67%	0.009%	$D_{US-30s-F}$	EER	2.95%	0.010%	$D_{US-60s-F}$	EER	1.88%	0.035%
	$D_{US-20s-M}$	FPR_1	99.76%	0.029%	$D_{US-30s-M}$	FPR_1	99.74%	0.008%	$D_{US-60s-M}$	FPR_1	99.90%	0.003%
	$D_{US-20s-F}$	FPR_1	99.61%	0.004%	$D_{US-30s-F}$	FPR_1	99.69%	0.004%	$D_{US-60s-F}$	FPR_1	99.79%	0.009%
D03	$D_{US-20s-M}$	FPR_2	99.66%	0.021%	$D_{US-30s-M}$	FPR_2	99.64%	0.058%	$D_{US-60s-M}$	FPR_2	99.87%	0.003%
	$D_{US-20s-F}$	FPR_2	99.52%	0.007%	$D_{US-30s-F}$	FPR_2	99.53%	0.006%	$D_{US-60s-F}$	FPR_2	99.74%	0.018%
	$D_{US-20s-M}$	FPR_3	99.86%	0.015%	$D_{US-30s-M}$	FPR_3	99.84%	0.011%	$D_{US-60s-M}$	FPR_3	99.95%	0.000%
	$D_{US-20s-F}$	FPR_3	99.79%	0.004%	$D_{US-30s-F}$	FPR_3	99.82%	0.006%	$D_{US-60s-F}$	FPR_3	99.87%	0.017%
	$D_{US-20s-M}$	EER	63.15%	0.033%	$D_{US-30s-M}$	EER	61.51%	0.052%	$D_{US-60s-M}$	EER	64.31%	0.009%
	$D_{US-20s-F}$	EER	60.19%	0.009%	$D_{US-30s-F}$	EER	60.49%	0.005%	$D_{US-60s-F}$	EER	62.29%	0.018%
	$D_{US-20s-M}$	FPR_1	74.93%	0.132%	$D_{US-30s-M}$	FPR_1	81.39%	0.278%	$D_{US-60s-M}$	FPR_1	76.29%	0.051%
	$D_{US-20s-F}$	FPR_1	52.45%	0.024%	$D_{US-30s-F}$	FPR_1	41.52%	0.022%	$D_{US-60s-F}$	FPR_1	64.68%	0.125%
D04	$D_{US-20s-M}$	FPR_2	75.93%	0.192%	$D_{US-30s-M}$	FPR_2	81.99%	0.248%	$D_{US-60s-M}$	FPR_2	77.25%	0.056%
	$D_{US-20s-F}$	FPR_2	53.62%	0.037%	$D_{US-30s-F}$	FPR_2	42.47%	0.051%	$D_{US-60s-F}$	FPR_2	66.08%	0.147%
	$D_{US-20s-M}$	FPR_3	74.16%	0.149%	$D_{US-30s-M}$	FPR_3	80.83%	0.240%	$D_{US-60s-M}$	FPR_3	75.33%	0.059%
	$D_{US-20s-F}$	FPR_3	51.36%	0.016%	$D_{US-30s-F}$	FPR_3	40.49%	0.050%	$D_{US-60s-F}$	FPR_3	63.55%	0.162%
	$D_{US-20s-M}$	EER	27.96%	0.086%	$D_{US-30s-M}$	EER	33.98%	0.236%	$D_{US-60s-M}$	EER	26.83%	0.027%
	$D_{US-20s-F}$	EER	23.02%	0.020%	$D_{US-30s-F}$	EER	19.18%	0.025%	$D_{US-60s-F}$	EER	25.85%	0.040%
	$D_{US-20s-M}$	FPR_1	100.00%	0.000%	$D_{US-30s-M}$	FPR_1	100.00%	0.000%	$D_{US-60s-M}$	FPR_1	100.00%	0.000%
	$D_{US-20s-F}$	FPR_1	100.00%	0.000%	$D_{US-30s-F}$	FPR_1	100.00%	0.000%	$D_{US-60s-F}$	FPR_1	100.00%	0.000%
D05	$D_{US-20s-M}$	FPR_2	100.00%	0.000%	$D_{US-30s-M}$	FPR_2	100.00%	0.000%	$D_{US-60s-M}$	FPR_2	100.00%	0.000%
	$D_{US-20s-F}$	FPR_2	100.00%	0.000%	$D_{US-30s-F}$	FPR_2	100.00%	0.000%	$D_{US-60s-F}$	FPR_2	100.00%	0.000%
	$D_{US-20s-M}$	FPR_3	100.00%	0.000%	$D_{US-30s-M}$	FPR_3	100.00%	0.000%	$D_{US-60s-M}$	FPR_3	100.00%	0.000%
	$D_{US-20s-F}$	FPR_3	100.00%	0.000%	$D_{US-30s-F}$	FPR_3	100.00%	0.000%	$D_{US-60s-F}$	FPR_3	100.00%	0.000%
	$D_{US-20s-M}$	EER	68.59%	0.072%	$D_{US-30s-M}$	EER	70.28%	0.113%	$D_{US-60s-M}$	EER	70.15%	0.024%
	$D_{US-20s-F}$	EER	66.56%	0.005%	$D_{US-30s-F}$	EER	69.57%	0.022%	$D_{US-60s-F}$	EER	72.71%	0.034%
	$D_{US-20s-M}$	FPR_1	85.37%	0.097%	$D_{US-30s-M}$	FPR_1	89.47%	0.166%	$D_{US-60s-M}$	FPR_1	72.08%	0.117%
	$D_{US-20s-F}$	FPR_1	85.78%	0.020%	$D_{US-30s-F}$	FPR_1	82.08%	0.065%	$D_{US-60s-F}$	FPR_1	80.66%	0.135%
D06	$D_{US-20s-M}$	FPR_2	80.23%	0.169%	$D_{US-30s-M}$	FPR_2	85.61%	0.237%	$D_{US-60s-M}$	FPR_2	65.61%	0.063%
	$D_{US-20s-F}$	FPR_2	80.10%	0.025%	$D_{US-30s-F}$	FPR_2	74.87%	0.021%	$D_{US-60s-F}$	FPR_2	71.35%	0.145%
	$D_{US-20s-M}$	FPR_3	90.53%	0.101%	$D_{US-30s-M}$	FPR_3	93.79%	0.166%	$D_{US-60s-M}$	FPR_3	79.90%	0.032%
	$D_{US-20s-F}$	FPR_3	91.44%	0.011%	$D_{US-30s-F}$	FPR_3	88.88%	0.039%	$D_{US-60s-F}$	FPR_3	88.94%	0.104%
	$D_{US-20s-M}$	EER	43.51%	0.112%	$D_{US-30s-M}$	EER	46.43%	0.135%	$D_{US-60s-M}$	EER	36.06%	0.062%
	$D_{US-20s-F}$	EER	42.12%	0.019%	$D_{US-30s-F}$	EER	39.81%	0.033%	$D_{US-60s-F}$	EER	34.96%	0.068%

4. Obtaining ΔFPR and ΔEER

In this section, we describe how we obtain **ΔFPR** and **ΔEER** reported in the paper from the absolute value of FPR and EER reported in Sec. 4. We will use example of age bias study for male speakers *i.e.* Tab. 5 to inform about our calculations. Notice there are 6 different age groups. For each metric, we first obtained the minimum value. For example, FPR_1 for detector Wav2Vec2 has the minimum value for age group 30s in Tab. 5. We refer to this value as $minFPR_1$ and then report $\Delta FPR_1 := FPR_1 - minFPR_1$. Note the minimum will be different for each metric and detector. We did this so that bias study and results are not dependent on individual detector performance and help

Table 5. Absolute performance of detectors in male age bias study.

Method		teens		20s		30s		40s		50s		60s	
		Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
TSSDNet	FPR_1	97.20%	0.077%	98.23%	0.052%	98.52%	0.161%	97.74%	0.051%	97.74%	0.051%	98.56%	0.106%
	FPR_2	99.95%	0.022%	99.96%	0.026%	99.99%	0.008%	99.94%	0.011%	99.93%	0.013%	99.92%	0.015%
	FPR_3	79.01%	0.169%	82.62%	0.390%	83.12%	0.317%	76.50%	0.199%	78.71%	0.116%	91.84%	0.301%
	EER	44.56%	0.151%	46.49%	0.332%	43.20%	0.151%	41.98%	0.336%	44.36%	0.126%	57.87%	0.130%
Wav2Vec2	FPR_1	33.59%	0.353%	27.01%	0.271%	21.71%	0.176%	26.03%	0.307%	27.01%	0.215%	28.61%	0.341%
	FPR_2	92.00%	0.332%	90.62%	0.324%	83.82%	0.322%	85.04%	0.300%	87.36%	0.183%	93.07%	0.097%
	FPR_3	4.09%	0.154%	2.63%	0.128%	2.35%	0.184%	4.03%	0.182%	3.23%	0.134%	1.47%	0.073%
	EER	5.07%	0.045%	3.79%	0.086%	3.68%	0.120%	4.90%	0.164%	4.37%	0.051%	2.97%	0.096%
Spec-ResNet	FPR_1	99.91%	0.017%	99.76%	0.064%	99.75%	0.045%	99.46%	0.060%	99.82%	0.024%	99.89%	0.028%
	FPR_2	99.86%	0.029%	99.67%	0.088%	99.65%	0.061%	99.31%	0.061%	99.73%	0.032%	99.88%	0.024%
	FPR_3	99.95%	0.024%	99.85%	0.029%	99.84%	0.038%	99.74%	0.042%	99.89%	0.006%	99.96%	0.017%
	EER	63.39%	0.105%	63.15%	0.072%	61.46%	0.130%	61.13%	0.122%	62.10%	0.051%	64.32%	0.051%
PS3DT	FPR_1	69.56%	0.547%	75.08%	0.551%	81.67%	0.298%	79.80%	0.493%	80.77%	0.237%	76.26%	0.180%
	FPR_2	70.33%	0.423%	75.91%	0.408%	82.20%	0.347%	80.82%	0.364%	81.76%	0.083%	77.48%	0.454%
	FPR_3	68.52%	0.292%	73.80%	0.789%	80.86%	0.297%	79.51%	0.205%	79.69%	0.171%	75.43%	0.202%
	EER	27.01%	0.173%	28.04%	0.110%	33.91%	0.186%	28.20%	0.242%	29.81%	0.082%	26.77%	0.073%
LFCC-GMMs	FPR_1	100.00%	0.000%	100.00%	0.000%	100.00%	0.000%	100.00%	0.000%	100.00%	0.000%	100.00%	0.000%
	FPR_2	100.00%	0.000%	100.00%	0.000%	100.00%	0.000%	100.00%	0.000%	100.00%	0.000%	100.00%	0.000%
	FPR_3	100.00%	0.000%	100.00%	0.000%	100.00%	0.000%	100.00%	0.000%	100.00%	0.000%	100.00%	0.000%
	EER	67.14%	0.117%	68.57%	0.178%	70.33%	0.059%	68.21%	0.122%	69.04%	0.074%	70.21%	0.117%
MFCC-ResNet	FPR_1	87.08%	0.270%	85.10%	0.299%	89.78%	0.399%	84.20%	0.256%	83.22%	0.120%	71.81%	0.316%
	FPR_2	81.88%	0.331%	80.14%	0.355%	85.36%	0.255%	78.78%	0.608%	76.96%	0.289%	65.66%	0.312%
	FPR_3	92.06%	0.231%	90.48%	0.225%	93.55%	0.169%	89.73%	0.356%	88.89%	0.180%	79.84%	0.232%
	EER	43.94%	0.122%	43.51%	0.263%	46.27%	0.084%	43.27%	0.168%	40.52%	0.174%	35.93%	0.139%

to capture difference in performance by a detector on one age group versus another. We use similar approach for all calculating ΔFPR_2, ΔFPR_3, and ΔEER.

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 Interspeech*, pp. 1008–1012, September 2019, Graz, Austria. 1, 3

[2] A. Nautsch, X. Wang, N. Evans, T. H. Kinnunen, V. Vestman, M. Todisco, H. Delgado, M. Sahidullah, J. Yamagishi, and K. A. Lee, "ASVspoof 2019: Spoofing Countermeasures for the Detection of Synthesized, Converted and Replayed Speech," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 2, pp. 252–265, February 2021. 1, 3

[3] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," pp. 2087–2091, September 2015, Dresden, Germany. 1

[4] M. Alzantot, Z. Wang, and M. B. Srivastava, "Deep Residual Neural Networks for Audio Spoofing Detection," *Proceedings of Interspeech*, pp. 1078–1082, September 2019, Graz, Austria. 1, 2, 3

[5] 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. 1

[6] A. K. S. Yadav, Z. Xiang, K. Bhagtani, P. Bestagini, S. Tubaro, and E. J. Delp, "Compression Robust Synthetic Speech Detection Using Patched Spectrogram Transformer," *arXiv:2402.14205*, February 2024. 2, 3

[7] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," *Proceedings of the International Conference on Learning Representations*, May 2019, New Orleans, LA. 2

[8] 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. 2, 3

[9] 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. 2, 3

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

[11] O. Tange, "Gnu parallel - the command-line power tool," *;login: The USENIX Magazine*, vol. 36, no. 1, pp. 42–47, Feb 2011. [Online]. Available: http://www.gnu.org/s/parallel 2

Table 6. Absolute performance of detectors in female age bias study.

Method		**teens**		**20s**		**30s**		**40s**		**50s**		**60s**	
		Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
TSSDNet	FPR_1	98.45%	0.088%	96.79%	0.241%	97.69%	0.055%	97.66%	0.025%	95.88%	0.063%	91.91%	0.345%
	FPR_2	99.91%	0.015%	99.61%	0.074%	99.88%	0.019%	99.80%	0.005%	99.84%	0.019%	99.93%	0.018%
	FPR_3	83.16%	0.230%	81.91%	0.334%	77.49%	0.245%	63.33%	0.061%	73.85%	0.298%	65.90%	0.245%
	EER	44.22%	0.227%	45.12%	0.213%	44.39%	0.119%	38.30%	0.035%	42.70%	0.121%	43.03%	0.069%
Wav2Vec2	FPR_1	35.85%	0.325%	29.53%	0.430%	23.79%	0.284%	30.43%	0.028%	23.24%	0.216%	17.37%	0.324%
	FPR_2	92.46%	0.178%	91.52%	0.128%	90.04%	0.265%	94.16%	0.012%	88.75%	0.224%	92.79%	0.207%
	FPR_3	12.78%	0.229%	2.34%	0.180%	1.38%	0.067%	1.87%	0.015%	2.70%	0.117%	0.63%	0.041%
	EER	11.53%	0.152%	3.62%	0.104%	2.88%	0.117%	3.21%	0.011%	3.97%	0.059%	1.92%	0.073%
Spec-ResNet	FPR_1	99.80%	0.037%	99.64%	0.090%	99.68%	0.040%	99.70%	0.005%	99.00%	0.063%	99.78%	0.045%
	FPR_2	99.70%	0.046%	99.53%	0.034%	99.50%	0.024%	99.62%	0.000%	98.80%	0.031%	99.76%	0.053%
	FPR_3	99.89%	0.013%	99.80%	0.029%	99.81%	0.027%	99.84%	0.000%	99.54%	0.071%	99.87%	0.027%
	EER	58.91%	0.069%	60.14%	0.063%	60.50%	0.099%	61.52%	0.017%	59.13%	0.142%	62.33%	0.075%
PS3DT	FPR_1	62.05%	0.212%	52.87%	0.299%	41.59%	0.312%	55.38%	0.015%	42.80%	0.175%	64.47%	0.505%
	FPR_2	62.58%	0.302%	53.60%	0.245%	42.29%	0.256%	56.77%	0.057%	44.17%	0.419%	65.95%	0.173%
	FPR_3	61.27%	0.394%	51.55%	0.551%	40.47%	0.330%	53.91%	0.055%	41.68%	0.588%	63.62%	0.285%
	EER	30.58%	0.177%	22.97%	0.127%	19.18%	0.170%	20.66%	0.015%	18.08%	0.027%	25.74%	0.140%
LFCC-GMMs	FPR_1	100.00%	0.000%	100.00%	0.000%	100.00%	0.000%	100.00%	0.000%	100.00%	0.000%	100.00%	0.000%
	FPR_2	100.00%	0.000%	100.00%	0.000%	100.00%	0.000%	100.00%	0.000%	100.00%	0.000%	100.00%	0.000%
	FPR_3	100.00%	0.000%	100.00%	0.000%	100.00%	0.000%	100.00%	0.000%	100.00%	0.000%	100.00%	0.000%
	EER	65.89%	0.058%	66.47%	0.072%	69.57%	0.141%	67.87%	0.012%	68.27%	0.104%	72.64%	0.117%
MFCC-ResNet	FPR_1	89.20%	0.091%	85.84%	0.210%	81.96%	0.284%	80.44%	0.012%	77.84%	0.355%	80.58%	0.359%
	FPR_2	84.37%	0.266%	80.16%	0.311%	75.04%	0.270%	73.22%	0.051%	70.16%	0.306%	70.98%	0.351%
	FPR_3	93.37%	0.081%	91.60%	0.248%	88.94%	0.195%	87.87%	0.036%	85.39%	0.088%	89.02%	0.204%
	EER	44.66%	0.102%	42.08%	0.194%	39.74%	0.110%	38.88%	0.017%	37.25%	0.167%	34.80%	0.272%

Table 7. Absolute performance of detectors in male accent bias study.

Method		**Canadian**		**US**		**British**		**Australian**		**South Asian**	
		Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
TSSDNet	FPR_1	98.61%	0.068%	98.29%	0.179%	98.19%	0.177%	97.90%	0.014%	98.63%	0.105%
	FPR_2	99.99%	0.006%	99.94%	0.010%	99.96%	0.018%	100.00%	0.000%	99.95%	0.010%
	FPR_3	84.20%	0.141%	82.80%	0.604%	78.04%	0.352%	80.11%	0.036%	87.68%	0.288%
	EER	48.31%	0.120%	46.72%	0.274%	43.10%	0.193%	48.51%	0.025%	52.27%	0.283%
Wav2Vec2	FPR_1	27.40%	0.211%	27.08%	0.450%	26.84%	0.230%	21.75%	0.032%	55.06%	0.455%
	FPR_2	91.04%	0.207%	90.68%	0.333%	90.91%	0.148%	88.13%	0.027%	95.45%	0.117%
	FPR_3	1.92%	0.046%	2.71%	0.236%	2.76%	0.224%	0.97%	0.007%	6.83%	0.108%
	EER	3.58%	0.023%	3.75%	0.156%	4.02%	0.173%	2.52%	0.013%	7.32%	0.111%
Spec-Resnet	FPR_1	99.87%	0.006%	99.73%	0.075%	99.84%	0.046%	99.88%	0.000%	99.75%	0.058%
	FPR_2	99.77%	0.015%	99.64%	0.061%	99.79%	0.053%	99.81%	0.000%	99.63%	0.075%
	FPR_3	99.92%	0.007%	99.86%	0.026%	99.91%	0.032%	99.90%	0.000%	99.86%	0.049%
	EER	63.92%	0.035%	63.09%	0.064%	63.39%	0.063%	63.79%	0.007%	62.13%	0.090%
PS3DT	FPR_1	71.78%	0.076%	75.12%	0.789%	83.29%	0.200%	81.76%	0.042%	77.87%	0.279%
	FPR_2	72.92%	0.181%	75.20%	0.258%	84.05%	0.175%	82.62%	0.040%	78.41%	0.400%
	FPR_3	70.78%	0.251%	74.10%	0.575%	82.52%	0.226%	81.07%	0.054%	76.46%	0.182%
	EER	27.57%	0.182%	27.96%	0.282%	28.97%	0.238%	30.55%	0.016%	30.06%	0.169%
LFCC-GMMs	FPR_1	100.00%	0.000%	100.00%	0.000%	100.00%	0.000%	100.00%	0.000%	100.00%	0.000%
	FPR_2	100.00%	0.000%	100.00%	0.000%	100.00%	0.000%	100.00%	0.000%	100.00%	0.000%
	FPR_3	100.00%	0.000%	100.00%	0.000%	100.00%	0.000%	100.00%	0.000%	100.00%	0.007%
	EER	68.52%	0.058%	68.68%	0.224%	67.74%	0.130%	69.95%	0.008%	63.47%	0.124%
MFCC-Resnet	FPR_1	87.37%	0.185%	85.46%	0.331%	90.31%	0.286%	85.10%	0.039%	90.40%	0.212%
	FPR_2	82.22%	0.270%	80.38%	0.485%	85.85%	0.294%	79.47%	0.051%	85.71%	0.287%
	FPR_3	92.34%	0.137%	90.24%	0.517%	94.50%	0.213%	90.72%	0.022%	94.42%	0.197%
	EER	44.94%	0.116%	43.57%	0.117%	46.02%	0.105%	42.87%	0.023%	45.96%	0.153%

Table 8. Absolute performance of detectors in female accent bias study.

Method		**Canadian**		**US**		**British**		**Australian**		**South Asian**	
		Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
TSSDNet	FPR_1	95.05%	0.034%	96.73%	0.156%	97.84%	0.062%	99.22%	0.049%	99.02%	0.088%
	FPR_2	99.25%	0.009%	99.58%	0.037%	99.90%	0.000%	99.98%	0.009%	99.89%	0.055%
	FPR_3	80.24%	0.077%	81.74%	0.308%	77.77%	0.215%	91.80%	0.107%	91.36%	0.286%
	EER	45.07%	0.062%	45.19%	0.289%	43.12%	0.027%	60.72%	0.135%	49.14%	0.181%
Wav2Vec2	FPR_1	21.50%	0.049%	29.28%	0.419%	33.89%	0.057%	61.50%	0.201%	73.76%	0.593%
	FPR_2	89.62%	0.074%	91.48%	0.169%	91.62%	0.073%	97.13%	0.069%	98.62%	0.130%
	FPR_3	0.91%	0.011%	2.31%	0.101%	4.67%	0.065%	2.97%	0.096%	27.53%	0.618%
	EER	2.31%	0.000%	3.63%	0.203%	5.52%	0.054%	4.70%	0.061%	18.61%	0.389%
Spec-Resnet	FPR_1	99.66%	0.009%	99.56%	0.049%	99.74%	0.011%	99.84%	0.020%	98.78%	0.052%
	FPR_2	99.52%	0.018%	99.49%	0.102%	99.61%	0.009%	99.84%	0.017%	98.58%	0.102%
	FPR_3	99.84%	0.009%	99.80%	0.050%	99.82%	0.009%	99.94%	0.009%	99.34%	0.116%
	EER	59.69%	0.021%	60.21%	0.128%	59.53%	0.052%	63.66%	0.040%	57.34%	0.298%
PS3DT	FPR_1	42.60%	0.082%	52.48%	0.466%	62.85%	0.211%	79.54%	0.117%	67.80%	0.554%
	FPR_2	43.29%	0.057%	53.50%	0.969%	63.81%	0.180%	80.36%	0.137%	69.09%	0.465%
	FPR_3	41.85%	0.062%	51.43%	0.463%	61.47%	0.057%	78.63%	0.266%	66.57%	0.454%
	EER	20.02%	0.025%	22.73%	0.211%	27.63%	0.075%	26.95%	0.040%	28.36%	0.333%
LFCC-GMMs	FPR_1	100.00%	0.000%	100.00%	0.000%	100.00%	0.000%	100.00%	0.000%	100.00%	0.000%
	FPR_2	100.00%	0.000%	100.00%	0.000%	100.00%	0.000%	100.00%	0.000%	100.00%	0.000%
	FPR_3	100.00%	0.000%	100.00%	0.000%	99.98%	0.000%	100.00%	0.000%	100.00%	0.000%
	EER	67.88%	0.020%	66.67%	0.166%	68.36%	0.082%	67.90%	0.058%	56.90%	0.085%
MFCC-Resnet	FPR_1	86.47%	0.014%	85.82%	0.475%	80.61%	0.148%	91.32%	0.072%	89.29%	0.188%
	FPR_2	80.60%	0.077%	80.01%	0.538%	74.37%	0.063%	85.58%	0.173%	83.87%	0.639%
	FPR_3	92.74%	0.049%	91.27%	0.223%	86.93%	0.106%	95.44%	0.087%	93.93%	0.309%
	EER	42.34%	0.039%	42.37%	0.189%	40.75%	0.063%	45.45%	0.138%	43.44%	0.269%