I CAN HEAR YOU: SELECTIVE ROBUST TRAINING FOR DEEPFAKE AUDIO DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in AI-generated voices have intensified the challenge of detecting deepfake audio, posing further risks for the spread of scams and disinformation. To tackle this issue, we establish the largest public voice dataset to date, named **DeepFakeVox-HQ**, comprising 1.3 million samples, including 270,000 high-quality deepfake samples from 14 diverse sources. Despite previously reported high accuracy, existing deepfake voice detectors struggle with our diversely collected dataset, and their detection success rates drop even further under realistic corruptions and adversarial attacks. We conduct a holistic investigation into factors that enhance model robustness and show that incorporating a diversified set of voice augmentations is beneficial. Moreover, we find that the best detection models often rely on high-frequency features, which are imperceptible to humans and can be easily manipulated by an attacker. To address this, we propose the **F-SAT**: **F**requency-**S**elective **A**dversarial **T**raining method focusing on high-frequency components. Empirical results demonstrate that our training dataset boosts baseline model performance (without robust training) by 33%, and our robust training further improves accuracy by 7.7% on clean samples and by 29.3% on corrupted and attacked samples, over the state-of-the-art RawNet3 model.

1 INTRODUCTION

AI-generated voices have become increasingly realistic due to larger datasets and enhanced model capacities (Ju et al., 2024; Neekhara et al., 2024), and they have been used in many important applications (Calahorra-Candao & Martín-de Hoyos, 2024). However, the success of AI-synthesized human voices poses significant security risks, including deepfake voice fraud and scams (Tak et al., 2021; Sun et al., 2023; Yang et al., 2024). A recent CNN report reveals a fraud in Hong Kong where a finance worker sent $25 million to scammers after a video call with a deepfake 'chief financial officer'. The voice was created by an AI model, highlighting the risk of such technology.

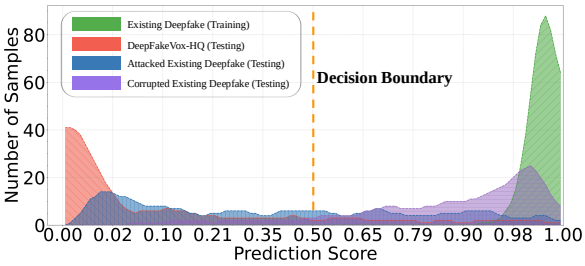


Figure 1: The distribution of deepfake samples over predicted scores using the state-of-the-art detector (Jung et al., 2022) trained on the In-the-Wild dataset (Müller et al., 2022), with a decision boundary at 0.5. Tests on original, corrupted, attacked, and real-world deepfake audio reveal significant shifts in prediction scores, highlighting that training solely on current public datasets without robust training methods leads to poor performance.

Due to the importance of this problem, a number of work has investigated detecting AI-generated audio. Despite previously reported high detection accuracy on public datasets (Todisco et al., 2019; Frank & Schönherr, 2021), existing deepfake voice detectors perform poorly under real-world conditions (Xu et al., 2020; Müller et al., 2022; Radford et al., 2023). This is because the established benchmarks are often trivial, small, outdated, and homogeneous. Consequently, models trained and validated solely on these datasets fail to generalize to more diverse and challenging real-world deepfake samples. Moreover, deep learning models for audio are particularly

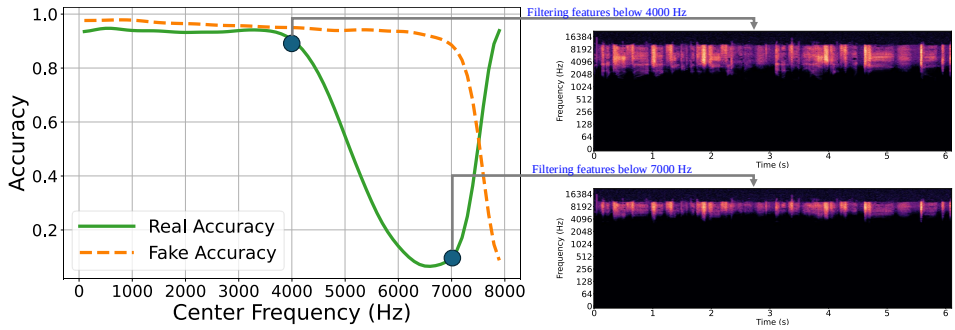


Figure 2: We apply a high-pass filter to audio samples to remove low-frequency components. The x-axis represents the center frequency of the filter applied. Notably, there is a marked decline in detection performance for real audio starting at 4000 Hz and for fake audio at 6000 Hz.

vulnerable to adversarial attacks (Szegedy et al., 2013) (Zhang et al., 2019), where an attacker can subtly alter audio inputs in ways that are imperceptible to humans but mislead models into incorrect classifications. Figure 1 illustrates a dramatic shift in the models' prediction scores when exposed to these factors, underscoring the need for more robust training methodologies.

To address the above limitations, we first created the largest deepfake audio dataset to date, *DeepFakeVox-HQ*, including 270,000 high-quality deepfake samples from 14 diverse and distinct origins. We show that simply training on our collected dataset can produce new state-of-the-art models.

Moreover, we find that even the state-of-the-art AI-voice detection models often depend on high-frequency features to make decisions (see Figure 2), which are imperceptible to humans. On the other hand, the low frequency signals can be heard by humans but are not relied on by the model to make predictions. As a result, natural corruptions in high frequency or attackers can easily manipulate the model by changing the high frequency signals, reducing the detection's robustness.

In an initial study, we observed that standard adversarial training on raw waveforms not only fails to bolster robustness but also diminishes performance on unattacked data. To address these shortcomings, we propose **F**requency-**S**elective **A**dversarial **T**raining (*F-SAT*), which focuses on high-frequency components. Since our adversarial training is targeted, we can mitigate specific vulnerabilities without touching the true features at lower frequencies, thus enhancing the model's resilience to corruptions and attacks while maintaining high accuracy on clean data.

Visualizations and empirical experiments demonstrate that using only our training dataset, we can produce state-of-the-art models, achieving a 33% improvement on the out-of-distribution portion of our test set, which includes 1,000 deepfake samples from the top five AI voice synthesis companies and 600 samples from social media. Additionally, by incorporating random audio augmentations, our model achieves the highest accuracy across 24 different types of corruptions. Furthermore, after applying F-SAT, our model further achieves a 30.4% improvement against adversarial attacks in the frequency domain and an 18.3% improvement against unseen attacks targeting raw waveform data in the time domain. We will release our data, code, and model upon acceptance.

2 Related Work

AI-synthesized human voice: AI voice synthesis generally falls into two categories: text-to-speech (TTS) and voice conversion (VC). TTS systems convert written text into spoken audio using a desired voice. This process typically involves three main components: a text analysis module that transforms text into linguistic features, an acoustic model that converts these features into a mel-spectrogram, and a vocoder. Some of the leading TTS models include StyleTTS (Li et al., 2024), VoiceCraft (Peng et al., 2024), XTTS (Casanova et al., 2024), and Tortoise-TTS (Betker, 2023), and are known for their ability to produce high-quality audio that closely mimics real human speech.

VC models, in contrast, take an audio sample from one person and transform it to sound like another person for the same speech content. Recent VC approaches primarily operate within the mel-spectrum domain (Ju et al., 2024; Shen et al., 2023; Popov et al., 2021), using deep neural networks to shift the mel-spectrograms from the source to the target voice.

Detection Method: AI deepfake detection models based on deep learning can be grouped into two main categories: those processing raw audio end-to-end and those analyzing spectrum. The first category includes models like RawNet2, which employs Sinc-Layers (Ravanelli & Bengio, 2018) to extract features directly from waveforms, and RawGAT-ST, which utilizes spectral and temporal sub-graphs (Tak et al., 2021). RawNet3 (Jung et al., 2022), which begins by using parameterized filterbanks (Zeghidour et al., 2018) to extract a time-frequency representation from the raw waveform and then is followed by three backbone blocks with residual connections, a structural approach that sets it apart from ECAPA-TDNN (Desplanques et al., 2020).These models process the audio data in its raw form to capture nuanced details directly impacting model performance.

The second category of AI deepfake detection models involves transforming raw audio into spectrograms for analysis. This process utilizes extracted features such as Mel Frequency Cepstral Coefficients (MFCCs) (Sahidullah & Saha, 2012), Constant Q Cepstral Coefficients (CQCCs) (Todisco et al., 2017), bit-rate (Borzì et al., 2022), and Linear Frequency Cepstral Coefficients (LFCCs) (Li et al., 2021). The analysis of these features is conducted using traditional machine learning methods like Gaussian Mixture Models (GMMs) (Todisco et al., 2019) or advanced neural networks such as Bidirectional Long Short-Term Memory (Bi-LSTM) (Akdeniz & Becerikli, 2021), ResNet (Alzantot et al., 2019), and Transformers (Zhang et al., 2021). These methods enable deeper and more intricate pattern recognition, enhancing the model's ability to identify and classify deepfake audio accurately.

Adversarial Attack: Neural networks are highly vulnerable to nearly imperceptible perturbations, known as adversarial attacks (Szegedy et al., 2013). Although initial studies focused on image models, speech tasks are similarly susceptible. Notably, adversarial attacks generated on spectrograms can deceive 2D CNN models and, when converted back to waveforms, can also effectively fool 1D CNN models (Koerich et al., 2020).

Adversarial training, initially developed for image processing systems (Madry et al., 2017; Mao et al., 2019), has been increasingly adapted to the audio domain (Chen et al., 2023), particularly to enhance the robustness of applications such as Automatic Speech Recognition (ASR) and speaker verification systems. Another defense method, smoothing, based on additive noise masking, has demonstrated great improvements in model robustness for these tasks (Olivier et al., 2021; Cohen et al., 2019). Additionally, a defensive strategy that employs diffusion models to counter adversarial audio attacks (Wu et al., 2023) effectively smooths out perturbations and prevents attackers from altering audio signals without significantly compromising signal quality.

While these techniques enhance robustness, they hurt the detection on unattacked audio, highlighting a trade-off between robustness and accuracy (Zhang et al., 2019; Tsipras et al., 2018). This compromise is particularly critical in scenarios that demand high accuracy and user satisfaction. Furthermore, the generalizability of adversarially trained models to new and unseen attacks remains limited (Rajaratnam et al., 2018), raising questions about their effectiveness in rapidly evolving threat environments.

3 DEEPFAKEVOX-HQ

3.1 TRAINING DATASET

In this section, we introduce a new training dataset and a rigorous test set. In contrast to prior dataset, our dataset is large, diversified, realistic, and up-to-date, as shown in Table **??**. Prior detectors show poor generalization capabilities in realistic settings, as shown in Figure 3. Both our training and testing datasets integrate the latest advancements in AI voice synthesis technologies. Additionally, the testing dataset includes several new models not covered in the training dataset, specifically designed to test the generalization ability of our detection systems.

High quality deepfake samples: The limitations of existing public datasets, which often lack high-quality deepfake samples, can potentially impair model performance. To address this, we have investigated more than 30 recent advancements in Text-to-Speech (TTS) and Voice Conversion (VC) models developed in the past few years. Our training dataset now includes deepfake audio samples generated using the top seven TTS models: MetaVoice-1B (Liu et al., 2021), StyleTTS-v2 (Li et al., 2024), VoiceCraft (Peng et al., 2024), WhisperSpeech (Radford et al., 2023), VokanTTS,

Dataset	#Real Utt	#Fake Utt	Language	Conditions	Year	#Fake Source	Fake Type
ASVSpoof19	12k	109k	English	Clean	2019	17	TTS, VC
ASVSpoof21	22k	589k	English	Clean, Noisy	2021	UNK	TTS, VC
WavFake	14k	90k	English, Japanese	Clean	2021	7	TTS
ADD2022	61k	251k	Chinese	Clean	2022	UNK	TTS, VC
In-The-Wild	20k	12k	English	Clean, Noisy	2022	19	TTS
LibriSeVoc	13k	79k	English	Clean	2023	6	TTS
Our Train	690k	640k	English	Clean, Noisy	2024	40	TTS, VC
Our Test	3k	3k	English	Clean, Noisy	2024	15	TTS, VC

Table 1: Comparison of Deepfake Audio Datasets

XTTS-v2 (Casanova et al., 2024), and Elevenlabs. We use four datasets—VCTK (Yamagishi, 2012), LibriSpeech (Panayotov et al., 2015), In-The-Wilds (Müller et al., 2022), and AudioSet (Gemmeke et al., 2017)—to generate deepfake audio. These datasets include both clean, high-quality and noisy, low-quality real audio, ensuring that the deepfake audio produced is highly diverse and accurately reflects real-world conditions.

Reference data: For both fake and real audio, having only high-quality samples is insufficient. A broader diversity of samples is essential for the training dataset. Thus, for real audio, we utilize portions from six public audio datasets: VCTK, LibriSpeech, AudioSet, ASRspoof2019 (Todisco et al., 2019), Voxceleb1 (Nagrani et al., 2017), and ASRspoof2021 (Liu et al., 2023), with half consisting of clean audio and the other half of noisy audio. For fake audio, we include two low-quality fake audio datasets: WaveFake (Frank & Schönherr, 2021) and ASRspoof2019 to further enhance the diversity of the training material.

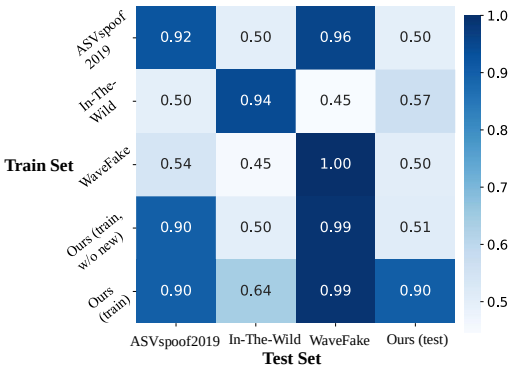


Figure 3: Performance of the RawNet3 baseline model on various datasets. 'Ours (train, w/o new)' represents our training dataset after removing all high-quality deepfake samples.

3.2 Testing Dataset

Our test dataset comprises approximately 7,000 samples, balanced equally between real and fake audio. Real audio samples are sourced from recent celebrity speeches and conversational videos on platforms such as YouTube. For the fake audio, we not only utilize samples created using the seven latest TTS models but have also expanded our dataset to include contributions from eight of the most advanced AI voice synthesis models or commercial software currently available, namely CosyVoice (Du et al., 2024), Resemble, Speechify, LOVO AI, Artlist, and Lipsynthesis. Additionally, this set includes fake audio directly collected from social media platforms like YouTube and X, further enriching the dataset with a diverse range of real-world scenarios. This comprehensive composition is strategically designed to rigorously test the generalization capabilities of our model.

Insight: Figure 3 shows that training with previous public datasets yields lower accuracy on ours. Additionally, removing high-quality deepfake samples from our training set significantly also reduces accuracy, highlighting their importance."

4 Method

In this section, we present our selective adversarial training approach, F-SAT. We also present a taxonomy of the most common corruptions and attacks in audio processing, which can be used for robust evaluation in realistic settings. Additionally, we discuss the implementation of Rand-Augmentation for audio to further enhance the robustness of our detection system.

4.1 F-SAT: Frequency-Selective Adversarial Training

Let $\mathbf{x}$ be a waveform input audio, and $\mathbf{y}$ be its ground-truth category label, To perform classification, neural networks commonly learn to predict the category $\hat{\mathbf{y}} = \mathcal{F}_\theta(\mathbf{x})$ by optimizing the cross-entropy $H(\hat{\mathbf{y}}, \mathbf{y})$ between the predictions and the ground truth. We use RawNet3 (Jung et al., 2022) that can

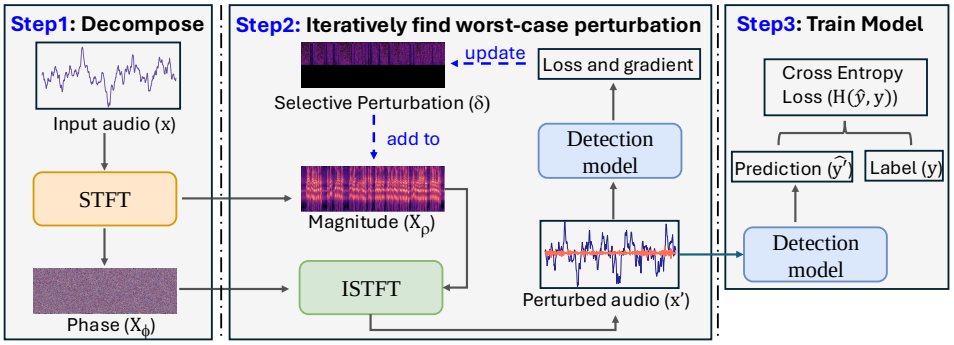


Figure 4: F-SAT Pipeline

process waveform audio input. The network parameters θ are estimated by minimizing the expected value of the objective:

$$\mathcal{L}_c(\mathbf{x}, \mathbf{y}) = H(F_\theta(\mathbf{x}), \mathbf{y}),$$

Time domain Attack: Adversarial attacks on audio can be directly added to the waveform. Let the additive perturbations be $\boldsymbol{\delta}$. For attacks in the time domain, we directly add $\boldsymbol{\delta}$ to the waveform $\mathbf{x}$. Due to the huge amount of freedom of the attack vector $\boldsymbol{\delta}$, the added attack vectors are often high frequency. We formulate this attack as: $\mathbf{x}' = \mathbf{x} + \boldsymbol{\delta}$.

Frequency Domain Attack: Attacks can also be applied in the frequency domain, accessed through a reversible Fourier transformation of the waveform. We transform the waveform $\mathbf{x}$ to the frequency domain $\mathbf{X}$ using the Short-Time Fourier Transform (STFT). The attack modifies $\mathbf{X}$ by adding $\boldsymbol{\delta}$, yielding $\mathbf{X}' = \mathbf{X} + \boldsymbol{\delta}$. The Inverse Short-Time Fourier Transform (ISTFT) then reverts $\mathbf{X}'$ back to the time domain, creating the manipulated audio waveform $\mathbf{x}'$.

Frequency-Selective Attack: Compared to time-domain attacks, frequency-domain attacks provide an inductive bias that enables the crafting of more controlled perturbations, such as those within a restricted bandwidth. Building on this, we introduce a Frequency-Selective Attack that targets specific frequency ranges within the magnitude component of a waveform. We define this attack as

$$\mathbf{x}' = s(\mathbf{x}, \boldsymbol{\delta}, f_l, f_u)$$

where f_l and f_u represent the lower and upper frequency boundaries of the range where the attack is to be applied. The procedure can be described as follows:

Starting with STFT result $\mathbf{X}$, we first compute its magnitude component $\mathbf{X}_\rho$ and phase component $\mathbf{X}_\phi$. To map the frequency range to the index range of the STFT spectrum, we calculate the lower and upper boundary indices r_l and r_u of the FFT spectrum using formulas:

$$r_l = \left\lfloor \frac{f_l \times n_{\text{fft}}}{sr} \right\rfloor, \quad r_u = \left\lceil \frac{f_u \times n_{\text{fft}}}{sr} \right\rceil,$$

where n_{fft} is the number of FFT points, and sr is the sampling rate of the original signal $\mathbf{x}$. We then define a mask function M using a diagonal matrix $\mathbf{D}$ such that

$$\mathbf{D}_{kk} = \begin{cases} 1 & \text{if } r_l \leq k \leq r_u \\ 0 & \text{otherwise} \end{cases}$$

This matrix $\mathbf{D}$ selectively targets the desired frequency components within the specified range $[r_l, r_u]$ for perturbation. The selective perturbation is then defined as:

$$\boldsymbol{\delta_s} = M(\boldsymbol{\delta}, r_l, r_u) = \mathbf{D} \cdot \boldsymbol{\delta}.$$

By adding this perturbation $\boldsymbol{\delta_s}$ to the magnitude component of the spectrogram, the perturbed spectrogram can be represented as:

$$\mathbf{X}' = (\mathbf{X}_\rho + \boldsymbol{\delta_s}) \cdot e^{j\mathbf{X}_\phi} = (\mathbf{X}_\rho + \mathbf{D} \cdot \boldsymbol{\delta}) \cdot e^{j\mathbf{X}_\phi}$$

Finally, we employ the ISTFT to convert the attacked spectrogram $\mathbf{X}'$ back into the time-domain signal $\mathbf{x}'$, thus generating the attacked audio waveform $\mathbf{x}'$.

Moreover, attackers design these worst-case perturbations to disrupt a trained network $\mathcal{F}_\theta$, aiming to maximize misclassifications by optimizing the objective:

$$\mathbf{x}' = \arg\max_{\mathbf{x}'} \mathcal{L}_c(\mathbf{x}', \mathbf{y}), \quad \text{s.t.} \quad \|\mathbf{x}' - \mathbf{x}\|_q < \epsilon,$$

where the perturbation $\boldsymbol{\delta} = \mathbf{x}' - \mathbf{x}$ is constrained by the q norm to be less than ϵ, ensuring minimal deviation.

F-SAT: Based on the Frequency-Selective Attack, we propose F-SAT, an adversarial training method which optimizes perturbations in the frequency domain by targeting the magnitude component within specific frequency ranges.

$$\mathbf{x}'_{n+1} = \Pi_{\mathbf{x}+S}\left(x'_n + \alpha \cdot M\left(\operatorname{sgn}\big(\nabla_{\mathbf{x}'}\mathcal{L}(s(\mathbf{x}, \boldsymbol{\delta}, f_l, f_u), y)\big), f_u, f_l\right)\right),$$

where $\Pi_{\mathbf{x}+S}$ represents the projection onto the allowable perturbation set S, defined by the condition $\|\mathbf{x}' - \mathbf{x}\|_p \leq \epsilon$. The parameter α denotes the step size of the update, and $\nabla_{\mathbf{x}'}\mathcal{L}$ is the gradient of the loss function with respect to the perturbed input $\mathbf{x}'$.

As illustrated in Figure 4, after K iterations (where $n \geq K$), we feed the most perturbed sample back into the detection model and calculate the cross-entropy loss between it and the ground truth label:

$$\mathcal{L}_{robust}(\mathbf{x}', \mathbf{y}) = H(F_\theta(\mathbf{x}'), \mathbf{y}),$$

Additionally, to maintain a balance between accuracy on original and attacked samples, we train the model using both the original and perturbed samples. We introduce a parameter $\boldsymbol{\gamma}$ to control the trade-off between clean loss and robust loss. The clean loss is defined as:

$$\mathcal{L}_{clean}(\mathbf{x}, \mathbf{y}) = H(F_\theta(\mathbf{x}), \mathbf{y}),$$

The total loss $\mathcal{L}_{total}$ is then computed as a weighted sum of clean and robust losses.

$$\mathcal{L}_{total} = \mathcal{L}_{clean} + \boldsymbol{\gamma} \cdot \mathcal{L}_{robust}$$

Since time domain attacks are often high frequency, by focusing the adversarial training on the high frequency, our approach not only enhances model robustness against frequency-targeted attacks, but also improves defenses against time-domain attacks.

4.2 RANDAUGMENT FOR AUDIO

Realistic Corruptions for Deepfake Audio Generation: As shown in Figure 9, we present a taxonomy for analyzing the robustness of deepfake audio detection systems that incorporate the most common corruptions and attacks. Corruptions in audio result from unintentional modifications during recording, processing, or transmission, impacting noise levels and frequency responses. Adversarial attacks, in contrast, are intentional manipulations aimed at deceiving detection systems through subtle changes across various attack methodologies.

RandAugment: Experiments show that the best detection model, RawNet3, experiences a great drop in accuracy when faced with audio corruptions. Inspired by the image-based RandAugment (Cubuk et al., 2020), which improves model robustness, we adapted this method for audio. We selectively apply $\mathcal{N}$ transformations from the available options, each assigned a uniform probability. An additional probability p determines whether each transformation is applied at a given instance, to balance the accuracy between original and corrupted samples. The magnitude of each transformation is controlled within predefined boundaries, with the intensity randomly sampled from this range.

5 EXPERIMENTS

We first compare our approach to existing state-of-the-art methods across three benchmarks and demonstrate improved accuracy. We then assess its robustness against corruption and adversarial attacks. We finally conduct ablation study on enhancements to the detection system's robustness.

5.1 DATASET

All models are trained on our training dataset and then evaluated on various benchmark datasets to test their generalization capabilities.

DeepFakeVox-HQ (test) Our custom test dataset incorporates 14 of the latest and highest quality TTS and VC models to generate fake audio. It also includes fake audio samples directly collected from social media platforms such as YouTube and X, providing a diverse set of real-world scenarios. For the experiments described in this paper, we specifically utilize seven fake sources not present in the training set. However, we also include samples from seven other fake sources used during training to facilitate future research.

ASVspoof2019 (Todisco et al., 2019) derived from the VCTK base corpus, which includes speech data captured from 107 speakers. It contains three major forms of spoofing attacks, namely synthetic, converted, and replayed speech.

WaveFake (Frank & Schönherr, 2021) Composed using six neural vocoders, this dataset features fake audio produced by MelGAN, Parallel WaveGAN, MB-MelGAN, FB-MelGAN, HiFi-GAN, and WaveGlow. WaveFake includes both English and Japanese audio. We only use the English part in this paper.

5.2 Baseline

RawNet2 (Tak et al., 2021) employs Sinc-Layers (Ravanelli & Bengio, 2018) to directly extract features from audio waveforms. These layers function as band-pass filters that enhance the detection of spoofed audio content.

RawGAT-ST (Tak et al., 2021) utilizes spectral and temporal sub-graphs integrated with a graph pooling strategy, effectively processing complex auditory environments.

TE-ResNet (Zhang et al., 2021) processes synthetic speech detection by first extracting MFCC from input speech. These coefficients are used as features for a CNN that extracts spatial features, followed by a Transformer that analyzes these to detect characteristics of synthetic speech effectively.

RawNet3 (Jung et al., 2022) begins by using parameterized filterbanks to extract a time-frequency representation from the raw waveform. This is followed by three backbone blocks with residual connections, a structural approach that deviates from ECAPA-TDNN (Desplanques et al., 2020).

5.3 Main Result

Results on Original Data: We trained all models on DeepFakeVox-HQ and tested them across three benchmarks: our test set, ASVspoof2019, and WaveFake, as shown in Table 2. Our method achieved state-of-the-art results across all three benchmarks. Compared with RawNet3, our method shows improvements of 7.7% points on DeepFakeVox-HQ (test), 8.4% points on ASVspoof2019, and 0.1% points on WaveFake.

Moreover, our baseline is divided into two categories: waveform-processing models, such as RawGAT-ST, RawNet2, and RawNet3, and the spectrum-processing TE-TesNet. Compared to the spectrum-based TE-TesNet, which only performs well on simpler datasets like WaveFake, the waveform models demonstrate superior effectiveness on our more diverse and complex dataset. This indicates that models with raw audio inputs are better equipped to capture detailed features in complex scenarios.

Figure 5 outlines results across various AI-voice synthesis models on our test set. For the fake sources are included in the training, all models demonstrated high accuracy . However, for unforeseen source, our method significantly outperformed others, particularly those from real-world social media platforms like YouTube and Lipsync, by up to 50% points, highlighting its superior generalization capabilities on out-of-distribution data.

Result on Corrupted data: Figure 6a shows detailed results for various corruptions. Our methods are depicted in the green region with F-SAT, and in the red region without F-SAT. They outperform all other baseline methods across 24 types of corruptions, with an average absolute increase of 15.3% points. However, as we can observe, aliasing has the most negative impact on accuracy. During aliasing, audio is intentionally downsampled and then resampled back, which not only removes high-frequency features but also introduces distortions characteristic of aliasing. In addition, low-pass filters, time stretching, and noise also significantly impact performance more than other tested corruptions.

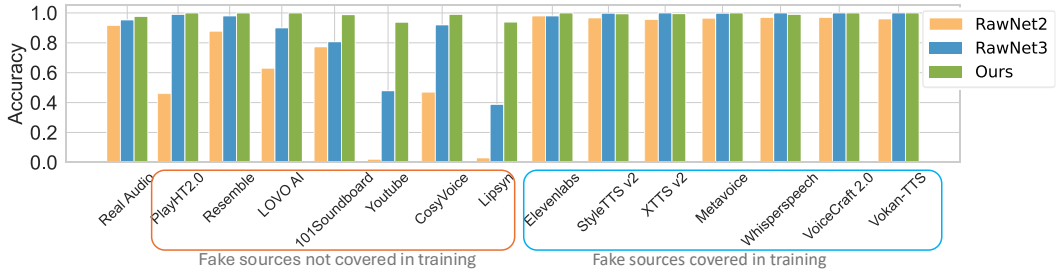


Figure 5: Detailed results across different AI-voice synthesis models on our test set. Real audio is directly sourced from recent YouTube content. Fake audio is divided into sources not included in training and those that are. Models evaluated include RawNet2, RawNet3, and Ours.

Approach	DeepFakeVox-HQ			ASVspoof2019			WaveFake		
	Real	Fake	Avg	Real	Fake	Avg	Real	Fake	Avg
TE-ResNet	85.9%	43.3%	64.6%	89.4%	97.3%	93.3%	96.9%	99.9%	98.4%
RawGAT-ST	92.6%	69.8%	81.2%	94.3%	95.1%	94.7%	98.2%	99.5%	98.8%
RawNet2	94.6%	57.8%	76.2%	93.0%	**98.5%**	95.8%	97.4%	100%	98.7%
RawNet3	96.1%	84.4%	90.3%	**97.0%**	90.8%	93.9%	99.9%	99.6%	99.8%
Ours	**97.5%**	**98.4%**	**98.0%**	95.0%	98.2%	**96.6%**	**99.9%**	**100%**	**99.9%**

Table 2: Comparative performance of our method and baseline models on our test set, ASVspoof2019, and WaveFake. (**Note:** WaveFake exclusively contains deepfake samples generated from LibriSpeech. We use a portion of LibriSpeech as the real sample for comparison. Results for our test set are calculated using only those deepfake samples whose sources were not covered during training.)

5.4 Results on Attacked Data

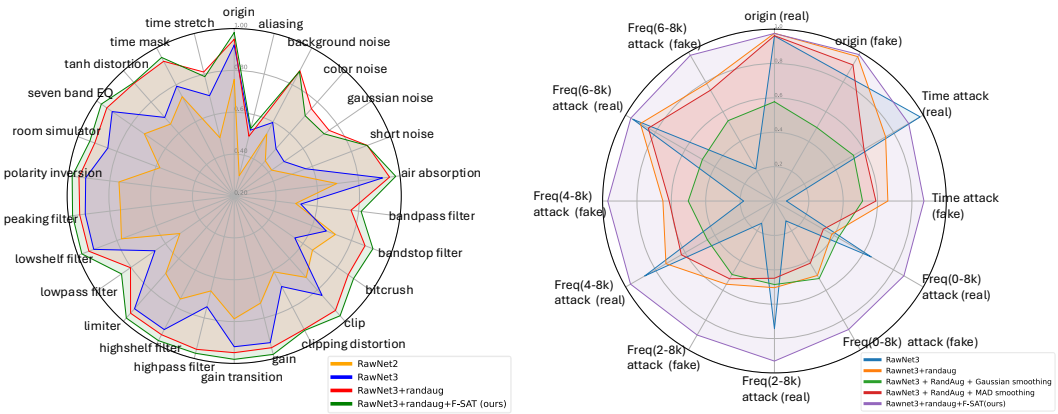
To assess our model's robustness against adversarial attacks, we employ the Projected Gradient Descent (PGD) attack, a well-established method in adversarial training. We compare our F-SAT with conventional adversarial training and smoothing defense methods.

Attack Settings for Evaluation and Perceptibility Assessment: In the main paper, we evaluate the l_∞-PGD attack separately on the time domain ($\epsilon = 1.00 \times 10^{-4}, \alpha = 4.00 \times 10^{-5}, iteration = 2$) and the magnitude component of the frequency domain($\epsilon = 1.00 \times 10^{-3}, \alpha = 4.00 \times 10^{-4}, iteration = 2$). For the frequency domain, we progressively expand the attacked frequency range to enlarge the attack surface. More comprehensive attack analyses are provided in Appendix A2. To ensure the imperceptibility of the attacks on both the time and frequency domains, we evaluate the Signal-to-Noise Ratio (SNR) of the attack compared to the original audio. The SNR values are 58.4 dB for the time domain attack and 68.7 dB for the frequency domain attack, indicating that the attacks remain imperceptible.

Compared with other defense method: We compare our Method F-SAT with other defense methods—MAD smoothing (Olivier et al., 2021), Gaussian smoothing (Cohen et al., 2019), and standard adversarial training (Mkadry et al., 2017)—on the RawNet3 model against various adversarial attacks in both time and frequency domains. As shown in Figure 6b, F-SAT outperforms all other methods.

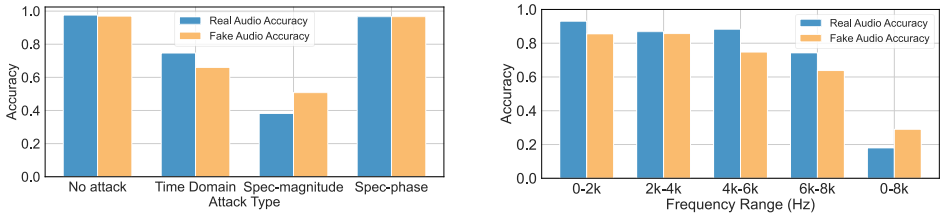
5.5 Ablation Study and Analysis

Choice of attack type for F-SAT: As discussed in Section 4.2, there are three types of adversarial attacks on audio: time domain, frequency domain on magnitude, and frequency domain on phase. All of these can be employed for adversarial training. However, our focus on targeting the magnitude component in attacks is explained in Figure 7a. Through our evaluation, we found that attacks based on magnitude in the frequency domain are particularly effective at degrading model performance. This is due to the fact that changes in magnitude directly affect the amplitude of audio signals, crucial for maintaining core acoustic features. Such changes alter the sound intensity across frequencies, complicating the model's task of distinguishing key characteristics between real and fake audio.



(a) Performance across various corruptions (b) Performance across different types of attacks

Figure 6: Comparative results showcasing the performance of our method against various types of corruptions and attacks. Our method outperforms the baseline across nearly all tested conditions.



(a) Impact of Different Attack Types (b) Frequency Range for Attack ($\epsilon = 0.01$)

Figure 7: Exploring Model Vulnerabilities: (a) shows accuracy under under various attack types; (b) details how attacks targeting magnitude in various frequency domains affect accuracy.

Conversely, phase attacks have minimal impact on the model's predictions, as phase alterations primarily influence spatial audio perception and do not significantly affect feature detection.

We further validate our conclusions experimentally, as shown in Table 3. Our findings indicate that F-SAT excels in adversarial training against attacks on both time and phase domains, especially for attacked samples. Surprisingly, incorporating adversarial training focused on the time domain appears to diminish overall accuracy on original data and does not effectively enhance robustness against attacks in either domain. Direct attacks in the time domain may disrupt most features critical for differentiating between real and fake audio, potentially overburdening the model during training.

Additionally, the introduction of RandAug to the RawNet3 baseline model significantly enhances accuracy on corrupted data, suggesting that the model benefits from exposure to varied conditions during training.

Approach	**Origin**		**Corruption**		**Attack(Time)**		**Attack(Frequency)**	
	Real	Fake	Real	Fake	Real	Fake	Real	Fake
RawNet3	96.1%	84.4%	94.7%	55.6%	**97.9%**	6.5%	80.4%	17.9%
RawNet3+RandAug	**97.6%**	97.0%	89.9%	**85.6%**	74.7%	66.0%	63.0%	62.4%
RawNet3+RandAug+AT(Time)	94.7%	83.1%	88.1%	78.2%	87.1%	12.4%	66.5%	16.5%
RawNet3+RandAug+ F-SAT	97.5%	**98.4%**	**94.8%**	85.0%	90.2%	**87.0%**	**93.3%**	**92.8%**

Table 3: Ablation study evaluating the impact of our specifically designed RandAug, F-SAT, and its comparative effects against time-domain adversarial training and phase-based adversarial training. (AT is an abbreviation of Adversarial Training)

Impact of frequency range for F-SAT: The importance of high-frequency components is underscored in Figure 7b, which shows the model's performance under gradient-based adversarial attacks across different frequency ranges. Attacks targeting higher frequencies notably degrade deepfake detection more than lower frequencies, underscoring the vulnerability of high-frequency features. By focusing attacks on high-frequency regions, we preserve low-frequency integrity, simplifying

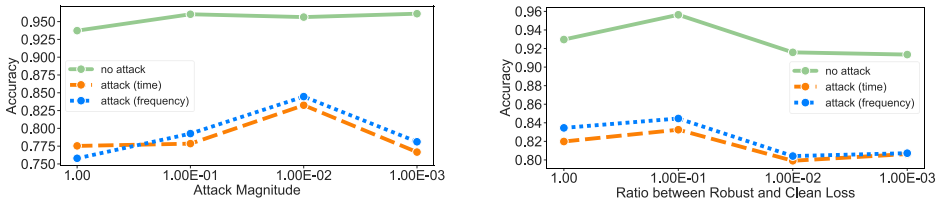
the model's task of distinguishing deepfake audio features. This approach maintains high accuracy on unattacked data and improves adversarial robustness.

Further analysis on selecting the optimal frequency range for F-SAT is presented in Table 4. Adversarial training within the 4k to 8k frequency range yields the best performance. We observe that increasing the attack frequency range decreases accuracy on original data due to the distortion of more critical features, adding complexity to the model's learning process.

Moreover, we observe that narrowing the frequency range of F-SAT increases the accuracy for attacked fake data while decreasing it for attacked real data. This supports the findings shown in Figure 2, which highlight a significant gap between the frequency domains where real and fake audio features predominantly exist. Fake audio features are concentrated in higher frequency ranges compared to real audio. In most real-world applications, criminals aim to make fake audio sound real enough to deceive detectors for committing fraud. Therefore, focusing adversarial training on high frequencies effectively enhances the robustness of fake audio.

Approach	**Original**		**Attack (Time)**		**Attack (0-8kHz)**		**Attack (2-8kHz)**		**Attack (4-8kHz)**		**Attack (6-8kHz)**	
	Real	Fake	Real	Fake	Real	Fake	Real	Fake	Real	Fake	Real	Fake
F-SAT (0-8kHz)	0.977	0.892	0.978	0.795	0.967	0.783	0.968	0.824	0.971	0.848	0.973	0.870
F-SAT (2-8kHz)	0.983	0.937	0.976	0.730	0.963	0.784	0.982	0.830	0.986	0.829	0.986	0.865
F-SAT (4-8kHz)	0.975	0.984	0.902	0.870	0.870	0.863	0.931	0.900	0.967	0.970	0.965	0.979
F-SAT (6-8kHz)	0.971	0.974	0.922	0.850	0.840	0.833	0.911	0.890	0.954	0.952	0.985	0.987

Table 4: Result for different range selection for F-SAT.



(a) Impact of Attack Magnitude ϵ on Detection Accuracy

(b) Impact of ratio γ between robust and clean loss (L_{robust}/L_{clean}) on Detection Accuracy

Figure 8: Exploration of hyperparameters to control attacks, balancing accuracy and robustness. (We train Rawnet3 + RandAug + F-SAT on 10% of DeepFakeVoc-HQ)

Balance Between Clean and Attacked Data: Figure 8a demonstrates that F-SAT achieves optimal performance with an attack magnitude of $\epsilon = 0.01$, maintaining high accuracy on both attacked and clean data. Figure 8b reveals that a ratio of 0.1 between robust and clean loss, (L_{robust}/L_{clean}), results in the highest overall accuracy. The results show that excessive focus on attacked data does not consistently boost robustness and may decrease clean data accuracy. Similarly, an undue emphasis on clean data does not reliably enhance accuracy and can weaken robustness. Thus, striking a balance between clean and attacked samples is critical for optimal model performance.

6 Conclusion

In our study, we introduce a dataset *DeepFakeVoc-HQ* that addresses diversity and quality issues in prior datasets, and provide a taxonomy to explore common audio corruptions and attacks. We find that leading AI voice detection models depend on vulnerable high-frequency features. This discovery leads us to develop F-SAT, a targeted adversarial training method that focuses on high-frequency components while while preserving the integrity of low-frequency features. Our approach effectively maintains accuracy on unattacked data and enhances robustness against various attacks. These results pioneer robust training for detecting fake audio for the first time, opening up a new direction for identifying such threats.

REFERENCES

Fulya Akdeniz and Yacsar Becerikli. Detection of copy-move forgery in audio signal with mel frequency and delta-mel frequency kepstrum coefficients. In *2021 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pp. 1–6. IEEE, 2021.

Moustafa Alzantot, Ziqi Wang, and Mani B Srivastava. Deep residual neural networks for audio spoofing detection. *arXiv preprint arXiv:1907.00501*, 2019.

James Betker. Better speech synthesis through scaling. *arXiv preprint arXiv:2305.07243*, 2023.

Stefano Borzì, Oliver Giudice, Filippo Stanco, and Dario Allegra. Is synthetic voice detection research going into the right direction? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 71–80, 2022.

Guillermo Calahorra-Candao and María José Martín-de Hoyos. The effect of anthropomorphism of virtual voice assistants on perceived safety as an antecedent to voice shopping. *Computers in Human Behavior*, 153:108124, 2024.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, et al. Xtts: a massively multilingual zero-shot text-to-speech model. *arXiv preprint arXiv:2406.04904*, 2024.

Guangke Chen, Zhe Zhao, Fu Song, Sen Chen, Lingling Fan, Feng Wang, and Jiashui Wang. Towards understanding and mitigating audio adversarial examples for speaker recognition. *IEEE Transactions on Dependable and Secure Computing*, 20(5):3970–3987, 2023. doi: 10.1109/TDSC.2022.3220673.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pp. 1310–1320. PMLR, 2019.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 702–703, 2020.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*, 2020.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, et al. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*, 2024.

Joel Frank and Lea Schönherr. Wavefake: A data set to facilitate audio deepfake detection. *arXiv preprint arXiv:2111.02813*, 2021.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 776–780. IEEE, 2017.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, et al. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *arXiv preprint arXiv:2403.03100*, 2024.

Jee-weon Jung, You Jin Kim, Hee-Soo Heo, Bong-Jin Lee, Youngki Kwon, and Joon Son Chung. Pushing the limits of raw waveform speaker recognition. *arXiv preprint arXiv:2203.08488*, 2022.

Karl Michel Koerich, Mohammad Esmailpour, Sajjad Abdoli, Alceu de S Britto, and Alessandro L Koerich. Cross-representation transferability of adversarial attacks: From spectrograms to audio waveforms. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7. IEEE, 2020.

Xu Li, Na Li, Chao Weng, Xunying Liu, Dan Su, Dong Yu, and Helen Meng. Replay and synthetic speech detection with res2net architecture. In *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 6354–6358. IEEE, 2021.

Yinghao Aaron Li, Cong Han, Vinay Raghavan, Gavin Mischler, and Nima Mesgarani. Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Songxiang Liu, Dan Su, and Dong Yu. Meta-voice: Fast few-shot style transfer for expressive voice cloning using meta learning. *arXiv preprint arXiv:2111.07218*, 2021.

Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, et al. Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2507–2522, 2023.

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

Aleksander M**k**adry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *stat*, 1050(9), 2017.

Chengzhi Mao, Ziyuan Zhong, Junfeng Yang, Carl Vondrick, and Baishakhi Ray. Metric learning for adversarial robustness. *Advances in neural information processing systems*, 32, 2019.

Nicolas M Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger. Does audio deepfake detection generalize? *arXiv preprint arXiv:2203.16263*, 2022.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017.

Paarth Neekhara, Shehzeen Hussain, Subhankar Ghosh, Jason Li, Rafael Valle, Rohan Badlani, and Boris Ginsburg. Improving robustness of llm-based speech synthesis by learning monotonic alignment. *arXiv preprint arXiv:2406.17957*, 2024.

Raphael Olivier, Bhiksha Raj, and Muhammad Shah. High-frequency adversarial defense for speech and audio. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2995–2999. IEEE, 2021.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210. IEEE, 2015.

Puyuan Peng, Po-Yao Huang, Daniel Li, Abdelrahman Mohamed, and David Harwath. Voicecraft: Zero-shot speech editing and text-to-speech in the wild. *arXiv preprint arXiv:2403.16973*, 2024.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, Mikhail Kudinov, and Jiansheng Wei. Diffusion-based voice conversion with fast maximum likelihood sampling scheme. *arXiv preprint arXiv:2109.13821*, 2021.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

Krishan Rajaratnam, Kunal Shah, and Jugal Kalita. Isolated and ensemble audio preprocessing methods for detecting adversarial examples against automatic speech recognition. *arXiv preprint arXiv:1809.04397*, 2018.

Mirco Ravanelli and Yoshua Bengio. Speaker recognition from raw waveform with sincnet. In *2018 IEEE spoken language technology workshop (SLT)*, pp. 1021–1028. IEEE, 2018.

Chandan KA Reddy, Vishak Gopal, and Ross Cutler. Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6493–6497. IEEE, 2021.

Md Sahidullah and Goutam Saha. Design, analysis and experimental evaluation of block based transformation in mfcc computation for speaker recognition. *Speech communication*, 54(4):543–565, 2012.

Kai Shen, Zeqian Ju, Xu Tan, Yanqing Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. *arXiv preprint arXiv:2304.09116*, 2023.

Chengzhe Sun, Shan Jia, Shuwei Hou, and Siwei Lyu. Ai-synthesized voice detection using neural vocoder artifacts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 904–912, 2023.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. End-to-end anti-spoofing with rawnet2. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6369–6373. IEEE, 2021.

Massimiliano Todisco, Héctor Delgado, and Nicholas Evans. Constant q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. *Computer Speech & Language*, 45: 516–535, 2017.

Massimiliano Todisco, Xin Wang, Ville Vestman, Md Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi Kinnunen, and Kong Aik Lee. Asvspoof 2019: Future horizons in spoofed and fake audio detection. *arXiv preprint arXiv:1904.05441*, 2019.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

Shutong Wu, Jiongxiao Wang, Wei Ping, Weili Nie, and Chaowei Xiao. Defending against adversarial audio via diffusion model. *arXiv preprint arXiv:2303.01507*, 2023.

Ruilin Xu, Rundi Wu, Yuko Ishiwaka, Carl Vondrick, and Changxi Zheng. Listening to sounds of silence for speech denoising. *Advances in Neural Information Processing Systems*, 33:9633–9648, 2020.

Junichi Yamagishi. English multi-speaker corpus for cstr voice cloning toolkit. *URL http://homepages. inf. ed. ac. uk/jyamagis/page3/page58/page58. html*, 2012.

Yujie Yang, Haochen Qin, Hang Zhou, Chengcheng Wang, Tianyu Guo, Kai Han, and Yunhe Wang. A robust audio deepfake detection system via multi-view feature. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 13131–13135. IEEE, 2024.

Neil Zeghidour, Nicolas Usunier, Iasonas Kokkinos, Thomas Schaiz, Gabriel Synnaeve, and Emmanuel Dupoux. Learning filterbanks from raw speech for phone recognition. In *2018 IEEE international conference on acoustics, speech and signal Processing (ICASSP)*, pp. 5509–5513. IEEE, 2018.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pp. 7472–7482. PMLR, 2019.

Zhenyu Zhang, Xiaowei Yi, and Xianfeng Zhao. Fake speech detection using residual network with transformer encoder. In *Proceedings of the 2021 ACM workshop on information hiding and multimedia security*, pp. 13–22, 2021.

A APPENDIX

A.1 EXPLANATION OF SOME TYPES OF CORRUPTIONS

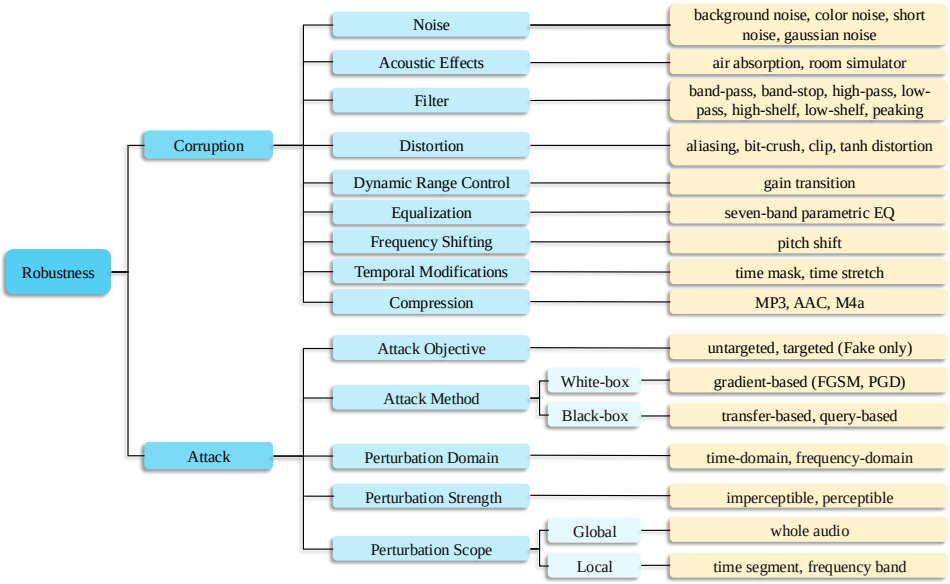


Figure 9: Overview of various corruption types and adversarial attack strategies affecting audio robustness. The diagram categorizes different forms of corruptions (e.g., noise, filtering, distortion) and adversarial attacks (e.g., white-box, black-box) based on their methods, objectives, and scope of perturbation. This framework outlines the challenges in ensuring the robustness of audio systems against both environmental corruption and intentional adversarial manipulation.

Air Absorption: Air absorption refers to the phenomenon where high-frequency sound waves are more strongly attenuated by air molecules. As sound travels through the air, it loses energy, particularly at higher frequencies, due to air viscosity and thermal conduction.

Room Simulator: A room simulator is a digital tool that emulates the acoustics of different rooms or spaces. It includes parameters such as room size, wall materials, and shape, which affect how sound reflects and diffuses.

Peaking: Peaking refers to adjusting a specific range of frequencies around a central frequency. It is commonly used in parametric equalizers.

Aliasing: Aliasing occurs when high-frequency components are sampled below the Nyquist rate, causing them to appear as lower frequencies. This can be represented using the Nyquist theorem:

$$f_{\text{sample}} > 2 \cdot f_{\text{max}},$$

where f_{sample} is the sampling frequency, and f_{max} is the maximum frequency of the signal. When this condition is violated, the signal is folded back into the lower frequencies, creating aliasing artifacts.

Bit-Crush: Bit-crushing reduces the bit depth of an audio signal, causing quantization errors.

Tanh Distortion: Tanh distortion is a form of soft clipping achieved using the hyperbolic tangent function. The output y is given by:

$$y = \tanh(kx),$$

where x is the input signal, and k controls the amount of distortion. As k increases, the function approximates a hard clipping effect.

Gain Transition: Gain transition refers to smoothly adjusting the amplitude over time.

Seven-Band Parametric EQ: A seven-band parametric EQ provides independent control over seven frequency bands. Each band can be adjusted using three parameters: center frequency (f_0),

gain (G), and bandwidth (Q). The overall transfer function is a combination of individual band filters:

$$H(f) = \prod_{i=1}^{7} H_i(f),$$

where $H_i(f)$ represents the frequency response of the i-th band.

Time Mask: Time masking involves temporarily silencing or removing a segment of audio.

A.2 RandAugment for Audio

```python
audio_transforms = [
    'background_noise', 'color_noise', 'short_noise', 'gaussian_noise', 'air_absorption', 'room_simulator', 'band_pass',
    'band_stop', 'high_pass', 'low_pass', 'high_shelf', 'low_shelf', 'peaking', 'aliasing', 'bit_crush', 'clip',
    'tanh_distortion', 'gain_transition', 'seven_band_parametric_EQ', 'pitch_shift', 'time_mask', 'time_stretch'
]
def rand_augment_audio(sample, N, p):
    """ Apply random augmentations to an audio sample.
    Args:
        sample: An audio sample    N: Number of augmentations to apply    p: Probability of applying each augmentation
    Returns:
        An augmented audio sample
    """
    operations = np.random.choice(audio_transforms, N)
    return operations(sample) if random.random() < p else sample
```

Figure 10: Python code for RandAugment for audio

Details of RandAugment for Audio are shown in Figure 10. In our experiments, we set $N = 1$ and $p = 0.9$.

A.3 Robustness to compression

To evaluate the robustness of our detection model to compression, we tested two lossy formats: MP3 and AAC. The evaluation utilized RawNet3 combined with RandAug and F-SAT. As shown in the table 5, both MP3 and AAC compression had minimal impact on detection accuracy.

Format	Real	Fake	Avg
Origin (90% wav + 10% mp3)	97.50%	98.40%	98.00%
MP3	97.50%	97.60%	97.60%
ACC	96.90%	98.60%	97.80%

Table 5: Detection Results for Compressed Audio Formats

A.4 Comprehensive Analysis of all Attack Types

Table 6 presents the detailed adversarial attack results for RawNet3 and our method under various conditions. It includes both white-box and black-box approaches and examines the impacts of attacks in both the time and frequency domains, detailing the different attack hyperparameters used.

Domain	Iters	Restart	Source	RawNet3		RawNet3+RandAug		RawNet3+RandAug+F-SAT	
				Real Acc	Fake Acc	Real Acc	Fake Acc	Real Acc	Fake Acc
No Attack	-	-	-	96.1%	84.4%	**97.6%**	97.0%	97.5%	**98.4%**
Time	2	1	A	**97.9%**	6.50%	74.7%	66.0%	90.2%	**87.0%**
	5	1	A	**90.6%**	2.10%	34.4%	42.4%	66.7%	**78.9%**
	5	2	A	**90.2%**	1.70%	33.2%	39.4%	65.4%	**77.5%**
	2	1	A'	**98.9%**	8.10%	79.8%	69.0%	92.9%	**88.1%**
	5	1	A'	**92.1%**	2.90%	43.7%	52.4%	74.4%	**81.1%**
	5	2	A'	**91.6%**	3.20%	42.1%	50.3%	73.2%	**80.5%**
	2	1	B	100%	17.3%	**98.7%**	91.7%	**98.7%**	**96.5%**
	5	1	B	100%	16.0%	**98.9%**	92.1%	**98.7%**	**96.3%**
	5	2	B	100%	16.2%	**98.9%**	92.1%	**98.7%**	**96.3%**
Frequency (0-8k Hz)	2	1	A	65.0%	13.3%	38.5%	50.0%	**87.0%**	**86.3%**
	5	1	A	23.8%	2.86%	7.46%	14.9%	**63.8%**	**66.8%**
	5	2	A	23.0%	2.70%	7.30%	15.2%	**63.7%**	**66.8%**
	2	1	A'	77.3%	14.9%	50.6%	60.6%	**90.2%**	**88.3%**
	5	1	A'	31.6%	4.92%	14.9%	27.6%	**72.7%**	**74.6%**
	5	2	A'	23.81%	2.38%	7.78%	16.03%	**63.81%**	**66.35%**
	2	1	B	99.84%	34.60%	98.89%	90.48%	**98.73%**	**96.83%**
	5	1	B	99.84%	29.52%	98.89%	89.84%	**98.73%**	**96.83%**
	5	2	B	**100.00%**	31.90%	98.89%	89.37%	98.73%	**96.67%**
Frequency (2-8k Hz)	2	1	A	74.3%	14.9%	50.4%	55.9%	**93.1%**	**90.0%**
	5	1	A	34.0%	4.29%	12.9%	22.9%	**83.5%**	**81.0%**
	5	2	A	33.0%	4.44%	12.7%	21.4%	**83.0%**	**81.7%**
	2	1	A'	84.0%	16.5%	59.7%	63.5%	**94.0%**	**91.3%**
	5	1	A'	43.8%	7.00%	23.5%	35.2%	**87.5%**	**85.6%**
	5	2	A'	33.49%	3.97%	12.86%	21.27%	**82.86%**	**80.79%**
	2	1	B	99.84%	33.33%	98.89%	89.68%	**98.57%**	**96.98%**
	5	1	B	**100.00%**	28.41%	98.89%	89.68%	98.73%	**95.87%**
	5	2	B	99.84%	30.95%	99.05%	89.68%	**98.57%**	**96.03%**
Frequency (4-8k Hz)	2	1	A	87.2%	17.9%	72.9%	64.9%	**96.7%**	**97.0%**
	5	1	A	54.9%	7.78%	35.4%	36.3%	**96.2%**	**95.4%**
	5	2	A	54.4%	6.83%	35.1%	36.3%	**95.7%**	**95.2%**
	2	1	A'	91.4%	18.3%	80.0%	71.9%	**97.1%**	**97.6%**
	5	1	A'	64.9%	10.5%	47.3%	50.6%	**96.5%**	**96.5%**
	5	2	A'	54.6%	7.1%	36.0%	36.3%	**96.0%**	**95.4%**
	2	1	B	99.7%	34.4%	99.0%	90.0%	**98.0%**	**98.3%**
	5	1	B	99.7%	29.2%	99.0%	89.5%	**98.0%**	**98.3%**
	5	2	B	99.7%	32.5%	99.2%	89.8%	**98.0%**	**98.3%**
Frequency (6-8k Hz)	2	1	A	95.2%	21.6%	90.1%	78.9%	**96.5%**	**97.9%**
	5	1	A	83.8%	16.0%	70.6%	67.6%	**95.7%**	**97.3%**
	5	2	A	83.7%	15.7%	70.5%	67.0%	**95.9%**	**97.9%**
	2	1	A'	95.7%	21.9%	92.7%	82.4%	**96.7%**	**97.9%**
	5	1	A'	86.8%	16.8%	78.4%	72.4%	**96.0%**	**97.9%**
	5	2	A'	83.2%	15.9%	70.5%	65.4%	**95.9%**	**97.8%**
	2	1	B	99.7%	38.3%	98.9%	91.1%	**97.0%**	**98.4%**
	5	1	B	99.7%	32.2%	98.9%	90.5%	**97.1%**	**98.4%**
	5	2	B	99.7%	36.8%	99.0%	90.5%	**97.3%**	**98.4%**

Table 6: Adversarial Attack Results on RawNet3 and Its Variants under Various Conditions. For the attack scenarios, we include both white-box and black-box approaches: A represents tests with the same model and identical weights, while A' indicates the same model but with different weights, and B denotes tests on a completely different model. For attacks in the time domain, we use $\boldsymbol{\epsilon} = 10^{-4}$ and $\boldsymbol{\alpha} = 4 \cdot 10^{-5}$. For attacks in the frequency domain, the parameters are $\boldsymbol{\epsilon} = 10^{-3}$ and $\boldsymbol{\alpha} = 4 \cdot 10^{-4}$.

A.5 DATASET DETAILS

Fake Source	Total Duration (Hours)	Audio Count	Mean Duration (Seconds)
VCTK	14.1	12.0k	4.2
Librispeech (Train)	961.1	281.2k	12.3
In-The-Wilds	14.6	9.3k	5.7
ASRspoof2019 (LA)	11.9	12.5k	3.4
Voxceleb1	340.4	148.6k	8.2
Audioset (Narration)	50.1	12.2k	14.8

Table 7: Summary of fake audio sources data

Fake Source	Total Duration (Hours)	Audio Count	Mean Duration (Seconds)
Metavoice	189.1	61.7k	11.0
StyleTTS-v2	186.6	61.6k	10.9
XTTS-v2	175.5	61.8k	10.2
VoiceCraft	119.9	59.4k	7.3
Whisperspeech	155.2	61.9k	9.0
Vokan-TTS	161.7	61.6k	9.4
Elevenlabs	3.3	3.2k	3.7
ASRspoof2019 (LA)	97.8	109.0k	3.2
Wavefake (English)	198.7	117.9k	6.1

Table 8: Summary of fake audio sources data

	VCTK Speaker_id: p244				In-the-Wild Speaker: Alan Watts			
Model	**Ovrl MOS**	**Sig MOS**	**Bak MOS**	**P808 MOS**	**Ovrl MOS**	**Sig MOS**	**Bak MOS**	**P808 MOS**
Real refer	3.26	3.56	4.04	3.61	3.02	3.40	3.74	3.57
metavoice	**3.29**	**3.58**	4.05	3.63	3.15	3.52	3.88	3.55
StyleTTS v2	3.28	3.56	**4.08**	**3.87**	**3.28**	**3.57**	**4.06**	**3.83**
XTTS v2	3.13	3.41	4.00	3.78	3.11	3.41	3.98	3.70
VoiceCraft	3.16	3.51	3.94	3.61	3.01	3.34	3.80	3.43
Whisperspeech	3.28	3.56	4.07	3.82	3.15	3.44	3.99	3.59
Vokan-TTS	3.23	3.55	4.01	3.71	2.94	3.39	3.66	3.60

Table 9: MOS Scores for Various TTS Models

Table 7 presents the real audio data we utilized from previously published datasets. For generating deepfake audio for the training set, we employ models such as XTTS v2, StyleTTS v2, Metavoice, Whisperspeech, Vokan-TTS, VoiceCraft, and Elevenlabs. Additionally, for the test set, we use Cosyvoice, PlayHT 2.0, Resemble, LOVO AI, and Lipsynthesis to create deepfake voices. We introduce post-processing augmentations to generate noisy deepfakes. Four real datasets—VCTK (12.0k), Librispeech-clean-100 (28.5k), Audioset (narration) (12.2k), and In-The-Wild (real parts: 9.3k)—are utilized to generate deepfake voices for the training set. Details of our generated deepfake audio and references to previous public deepfake audio datasets are presented in Table 8. Additionally, we employed DNSMOS (Reddy et al., 2021) to quantitatively measure the synthetic speech quality across these models, using a scale from 1 to 5, where higher values indicate better quality. As demonstrated in Table 9, Metavoice and StyleTTS outperform other AI voice synthesis models.

A.6 F-SAT'S TRAINING EFFICIENCY

F-SAT's training efficiency is influenced by hyperparameters such as attack iterations and restart counts, which identify the worst-case perturbation. Drawing on insights from "Fast Is Better Than Free: Revisiting Adversarial Training," we optimized these parameters by setting restarts to one and attack iterations to one or two, while employing a larger attack magnitude to enhance robustness. Training time is shown in the Table 10. Although F-SAT requires longer training times, it improves accuracy by an average of 9% on original data and 43% on attacked data compared to Standard Adversarial Training. We should not compromise accuracy merely to accelerate training.

Description	w/o Adversarial Training	Standard Adversarial Training	F-SAT
Training Duration (Days)	2	4.5	8
Number of Epochs	15	15	15
Hardware Used	A100 GPU	A100 GPU	A100 GPU

Table 10: Training time comparasion

A.6.1 TRAINING HYPERPARAMETER

Here are the Training hyperparameter of F-SAT for Table 3:

Training Hyperparameters

- **Learning Rate (lr):** 1×10^{-5}
- **Epochs:** 15
- **Batch Size (bs):** 16
- **Optimizer:** adam
- **Augmentation Number (aug_num):** 1 or 2
- **Augmentation Probability (aug_prob):** 0.9

LR Scheduler (Warmup Cosine)

- **Warm-up Epochs:** 1
- **Warm-up LR:** 1×10^{-6}
- **Minimum LR:** 1×10^{-7}

Attack Hyperparameters

- **Attack Type:** l_∞
- **Epsilon:** 0.005, **Alpha:** 0.002
- **Gamma (control ratio of clean loss and robust loss):** 0.1
- **Attack Iterations:** 2
- **Restarts:** 1

Mixup Hyperparameters

- **Mixup Alpha:** 0.5

A.7 GENERALIZATION TO SPECTURM-BASED MODEL

We further investigate whether our method can be generalized to spectrum-based models, despite its relatively inferior performance compared to models utilizing raw waveform inputs. We selected a prior baseline model, TE-ResNet, for evaluation. Unlike models trained on raw waveforms, TE-ResNet exhibits a different vulnerability pattern, with its susceptible features primarily concentrated in the lower frequency ranges. To address this, we applied targeted F-SAT on the 0-4kHz frequency range. The results, as shown in Table 11, indicate that incorporating RandAug into TE-ResNet improves model accuracy on the original data by 6.9%, while the additional F-SAT boosts overall robustness under attack scenarios, enhancing Attack Overall accuracy by 7.7%. These findings suggest that our method can be adapted to spectrum-based models, yielding enhanced performance and robustness.

Condition	**TE-ResNet**			**TE-ResNet + RandAug**			**TE-ResNet + SFAT**		
	Real	Fake	Avg	Real	Fake	Avg	Real	Fake	Avg
Original	85.9%	43.3%	64.6%	86.8%	56.2%	**71.5%**	52.4%	72.1%	62.3%
Attack Overall	94.3%	16.1%	55.2%	80.0%	19.7%	49.8%	74.6%	51.3%	**62.9%**
Attack (0-2000Hz)	97.9%	0.0%	49.0%	77.1%	10.0%	43.6%	91.6%	37.5%	**64.5%**
Attack (2000-4000Hz)	95.7%	13.1%	54.4%	81.8%	18.3%	50.0%	73.9%	51.4%	**62.7%**
Attack (4000-6000Hz)	92.7%	23.3%	58.0%	80.9%	23.9%	52.4%	68.6%	56.5%	**62.5%**
Attack (6000-8000Hz)	91.0%	28.1%	59.5%	80.3%	26.6%	53.4%	64.4%	59.8%	**62.1%**

Table 11: Results of the TE-ResNet model on our dataset, enhanced with RandAug and F-SAT. This table demonstrates how our method adapts to models with spectrum input, showcasing improvements and robustness.