IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 4 August 2023, accepted 19 August 2023, date of publication 31 August 2023, date of current version 7 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3310809

RESEARCH ARTICLE

On the Defense of Spoofing Countermeasures Against Adversarial Attacks

LONG NGUYEN-VU, (Member, IEEE), THIEN-PHUC DOAN, (Student Member, IEEE), MAI BUI, KIHUN HONG, (Member, IEEE), AND SOUHWAN JUNG, (Member, IEEE)

School of Electronic Engineering, Soongsil University, Seoul 06978, South Korea

Corresponding author: Kihun Hong (khong@ssu.ac.kr)

This work was supported in part by the Ministry of Science and ICT (MSIT), South Korea, through the Information Technology Research Center Support Program supervised by the Institute for Information and Communications Technology Planning and Evaluation (IITP) under Grant IITP-2023-2020-0-01602; and in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant funded by the Korea Government through MSIT (Robust deepfake audio detection development against adversarial attacks) under Grant RS-2023-00263037.

ABSTRACT Advances in speech synthesis have exposed the vulnerability of spoofing countermeasure (CM) systems. Adversarial attacks exacerbate this problem, mainly due to the reliance of most CM models on deep neural networks. While research on adversarial attacks in anti-spoofing systems has received considerable attention, there is a relative scarcity of studies focused on developing effective defense techniques. In this study, we propose a defense strategy against such attacks by augmenting training data with frequency band-pass filtering and denoising. Our approach aims to limit the impact of perturbation, thereby reducing the susceptibility to adversarial samples. Furthermore, our findings reveal that the use of Max-Feature-Map (MFM) and frequency band-pass filtering provides additional benefits in suppressing different noise types. To empirically validate this hypothesis, we conduct tests on different CM models using adversarial samples derived from the ASVspoof challenge and other well-known datasets. The evaluation results show that such defense mechanisms can potentially enhance the performance of spoofing countermeasure systems.

INDEX TERMS Automatic speaker verification, adversarial attack, spoofing countermeasure, psychoacoustics.

I. INTRODUCTION

The recent advancements in state-of-the-art generative models have revolutionized speech synthesis, enabling the creation of high-quality synthetic speech that closely emulates genuine speakers. Despite the advantages synthetic speech offers, the rise of audio deepfakes presents a severe threat to our society and economy. Deepfake audio employs advanced deep learning techniques like Text-to-Speech (TTS) or Voice Conversion (VC) to generate sophisticated spoofed audio samples, which can be considered presentation attacks [1]. Such maliciously crafted artificial speech can deceive both Automatic speaker verification systems (ASV), posing a severe threat to the integrity and reliability of speaker verification processes. Moreover, criminals can exploit speech samples generated by deep neural networks to conduct social engineering attacks or disseminate misinformation against human listeners. To evade detection, perpetrators may employ sophisticated techniques, including background noise, reverberation, or the creation of partially fake audio, all of which enhance the deception rate of the spoofing attempts.

In light of this emerging threat, developing robust countermeasure systems (CM) to defend against such spoofing attacks is crucial. Recent innovations in deep learning and end-to-end solutions offer promising avenues for combating voice spoofing attacks. Recently, several studies have focused on deep learning and end-to-end solutions, aiming to develop unified approaches to tackle various attacks. However, the exploration of such solutions is still in its early stages, emphasizing the urgent need for a comprehensive and unified approach in combating voice spoofing attacks in ASV systems [2], [3].

The associate editor coordinating the review of this manuscript and approving it for publication was Mohamad Afendee Mohamed.

In recent years, adversarial attacks have become a concerning problem in machine learning security. Several studies have discussed this topic in Automatic Speech Recognition (ASR) [4], [5], [6], [7], [8], [9] and in Automatic Speaker Verification/Spoofing Countermeasure (ASV-CM) [10], [11], [12], [13], [14], [15], [16], [17] tandem systems. While advances in speech synthesis allow a fake audio sample to sound indistinguishable from a genuine sample, adversarial attacks make it even more stealthy. Due to the transferability nature across deep learning-based models [18], [19], adversarial samples are often retrieved from a surrogate (substitute) model to attack the target (supposedly similar) model. As announced by the ASVspoof organizers, ASVspoof5 competition[1] will shift the focus to VC and TTS, including adversarial attacks relying on ASV/CM feedback. The need to build resilient defense mechanisms has never been more urgent, and addressing these challenges requires our utmost attention.

Among the defense mechanisms, adversarial training is a popular choice against adversarial attacks. This approach exposes the model to adversarial examples during training and forces it to make more robust predictions. Despite the advantage, adversarial training can be computationally expensive while generating less representative samples. This can lead to reduced performance on real-world data that does not conform to those specific attack types. Another approach was spatial smoothing derived from computer vision domain, which is an alternative technique to soften the audio signal features using the median and mean filters [11], [12] to restrict the capability of generated adversarial samples. However, its effectiveness depends on the size and shape of the smoothing filter, i.e., a trade-off between robustness and accuracy needs to be carefully balanced. In this study, we aim to defend against adversarial attacks on spoofing countermeasure systems (CM). We propose two techniques to improve the robustness of CM models by training them with augmented data resulting from band-pass filters and denoising.

Our hypotheses are as follows: (1) An original speech waveform would be too permissible for generating adversarial examples by making small changes to different waveform frequency ranges. Gradient-based attacks can perturb higher or lower frequency ranges without causing perceivable distortion to the auditory system. Therefore, a model trained with all frequencies may be more vulnerable to such attacks. (2) The countermeasure systems do not need to use a full range of frequencies because their goal is to identify spoofing samples. This distinction sets them apart from speech recognition, which requires audio samples to retain properties like intelligibility, consistency, naturalness, and emotion. (3) Spoofed speech samples are expected to exhibit similarities with bona fide samples in the "normal conversation" frequency range (refer to Fig. 1). It becomes essential for the CM to prioritize learning from the discriminative features present in this range. (4) Both band-pass filter and denoising techniques have the effect of reducing noise. We would like to see the potential of these techniques under adversarial scenarios.

[1]https://www.asvspoof.org

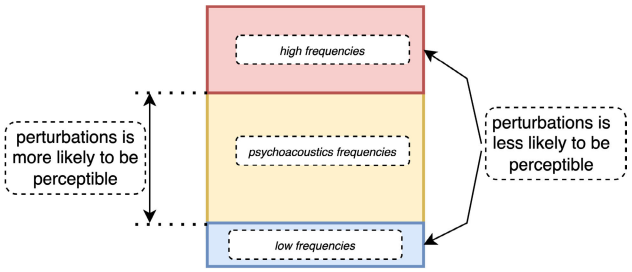


FIGURE 1. Perturbations in high and low frequencies beyond psychoacoustics range will be less likely to be perceptible.

In fact, previous attempts such as Wave-Guard [20] adopted transformation functions using low-shelf and high-shelf filters to remove adversarial noise from a given signal. Unlike our study, Wave-Guard did not completely remove the frequencies beyond some thresholds but lowered their amplitude, which is more suitable for Automatic Speaker Recognition (ASR) systems than CM systems. Motivated by such research, we explored practical neural network architectures such as LCNN and augmentation techniques to enhance the capability of countermeasure systems.

In this study, we made the following contributions:

- We propose two augmentation methods taking into account adversarial noise and perceptible frequencies.
- On the augmented datasets, we trained two types of neural networks commonly used in spoofing countermeasures, SENet and LCNN. The evaluation results show that models trained on the augmented datasets are more robust to adversarial samples than those trained with the original dataset. We also compared our enhanced model with other state-of-the-art models under various black-box datasets.
- We compared the band-pass filter technique with the denoising approach. The first is proven to be a stronger defense and can be a good candidate against psychoacoustics-based adversarial attacks.
- We conducted a rigorous evaluation to assess the efficacy of our proposed defense strategies, employing diverse datasets and state-of-the-art models. The results showcased the potential of our techniques in establishing universally robust models capable of withstanding various types of adversarial attacks.

Next, we discuss some background and related work in Section II. Our proposed scheme is detailed in Sections III and IV. We conclude our work in Section V. The source code required for conducting the experiments can be found at https://github.com/nguyenvulong/AdvDefenseCM.

II. BACKGROUND AND RELATED WORK

A. ADVERSARIAL ATTACKS

We briefly explain the two popular adversarial attacks we used in this work from the perspective of an audio

waveform. The FGSM attack [21] involves adding a small perturbation to the input data that maximizes the loss function (i.e., increasing the false acceptance rate of a CM). The perturbation is computed by taking the sign of the gradient of the loss function with respect to the input data.

$$X_{adv} = X + \epsilon \text{sign}(\nabla_X J(\theta, X, Y)) \quad (1)$$

where X is the original input data; X_{adv} is the processed audio input; Y is the ground-truth label; $J(\theta, X, Y)$ is the loss function parameterized by θ; ϵ is the magnitude of the perturbation (i.e., the size of the attack); $\nabla_X J(\theta, X, Y)$ is the gradient of the loss function with respect to the input X.

The PGD attack [22] is a more advanced adversarial attack that iteratively applies the FGSM with a smaller perturbation size, and projects the result back onto a valid data range.

$$X_{t+1} = \text{Clip}_X\{X_t + \alpha \text{sign}(\nabla_X J(\theta, X_t, Y))\} \quad (2)$$

where iteration step t and the step size α are introduced; $\text{Clip}_X(\cdot)$ is the clipping function to ensure that the X_{adv} is valid, i.e., no perceivable distortion.

B. ADVERSARIAL ATTACKS ON SPOOFING COUNTERMEASURES

Adversarial attacks can be categorized into white-box and black-box settings, depending on the knowledge of the attack about the target model [23]. Adversarial samples can be retrieved by launching adversarial attacks on a model while having full knowledge about it (i.e., white-box setting). Such samples can then be used to attack other models without knowing much about their architectures (hence the term black-box setting). In practical scenarios, a surrogate model is often used to indirectly launch a black-box attack [10], [24], [25]. The authors of [10] have shown that all models are vulnerable to FGSM and PGD black-box attacks, and smaller models are more vulnerable than larger ones to adversarial samples regarding the transferability effect. The attacks became even more severe when ensemble learning was applied [25]. Reference [26] performed the first targeted, over-telephony-network attack on the countermeasure (CM), which posed a serious threat in the emerging use of call centers.

Meanwhile, the defense can be grouped into two main categories: passive and active. Authors from [11], [12], and [14] proposed different passive defense methods to enhance the adversarial robustness of the system: spatial smoothing [11], layer-wise noise-to-signal ratio (LNSR) for robustness quantization [12], or cascaded models to purify the input samples [14]. The common characteristic among these techniques is that the input features have been transformed, thus alleviating the effect of the attack. Regarding active defense, adversarial training [11] is the popular choice besides adversarial sample detection approaches [13], [17], where the neural networks were specifically trained to spot these samples.

We summarized the studies of adversarial attacks on CM systems in Table 1.

C. FREQUENCY MASKING

Frequency masking is a similar technique used for training CM models [27], [28], [29], [30]. Specifically, [27] randomly dropped out a frequency band range during training as a means of data augmentation. RawGAT-ST [28] randomly selected contiguous *sinc* channels. Raw PC-DARTS [29] proposed *filter masking* as a regularizer. The authors of [30] proposed using Frequency Feature Masking (FFM) on five mel-spectrogram-based neural network architectures and showed that FFM is useful when noise may present in the high or low-frequency band. Different from this study, frequency masking has yet to be used for training a robust model against adversarial samples.

III. PROPOSED SCHEME

A. THREAT MODEL

Depending on the capability of an attacker, perturbations can prioritize high and low frequencies to avoid perceivable distortion because the quality of an audio file may degrade significantly to a human listener when perturbations are added in psychoacoustics frequencies (see Fig. 1). In this work, we assume that the adversarial attacks may happen in any frequency range but hypothesize that such attacks can be restricted in a narrower band after band-pass filtered, as depicted in Figure 3. In fact, the authors of [6], [31], and [32] discussed that psychoacoustics-based adversarial attacks are likely to exploit the properties of deep neural network-based CMs when trained with a wide band of frequency. For instance, [6] manipulated the acoustic signal below the human perception's threshold, whereas [32] found that an audio sample can be altered at lower frequencies without causing perceptible distortion. In fact, such assumption can also be generalized to other contexts beyond the audio domain. For instance, the authors of [33] sought to enhance the transferability of an adversarial attack by focusing on a low-frequency component of point clouds. The authors combined losses from the original point cloud and its low-frequency component to generate adversarial samples and demonstrated that the attack is more effective against state-of-the-art 3D defense methods.

B. SEARCH FOR OPTIMAL FREQUENCY RANGES

It is essential to determine suitable frequencies for training CM models, therefore we started by investigating the range that most impacts the auditory system. How humans perceive sound has been well-studied in the past. The frequency range between 20 Hz to 20,000 Hz is audible to the human auditory system, and the highest sensitivity is between 500 Hz and 4,000 Hz [34]. For everyday conversation, the range typically takes place between 500 Hz to 3,000 Hz [34].

While analyzing audio samples from the ASVspoof 2019 [35], we found that frequencies from 100Hz to 5kHz are sufficient for representing normal speech. Furthermore, we realized certain types of background noise can be removed in such frequencies, i.e., making the audio samples sound

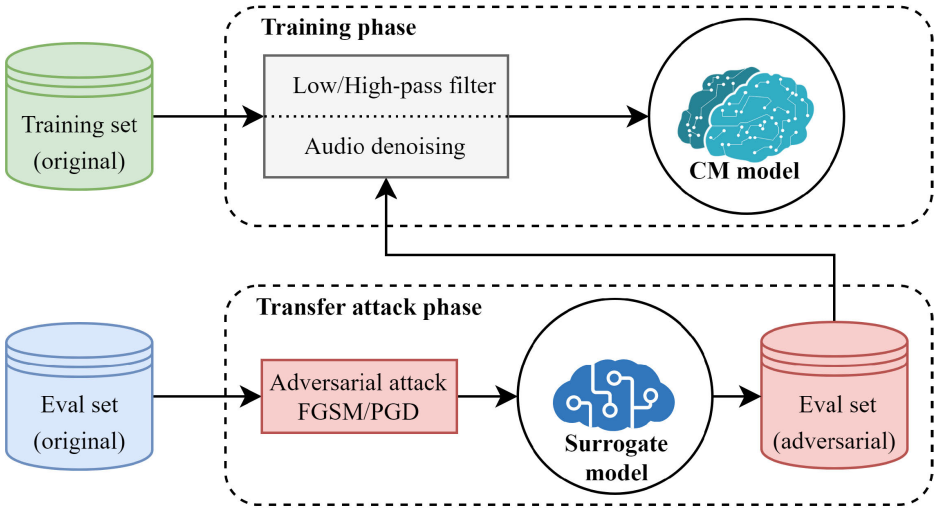
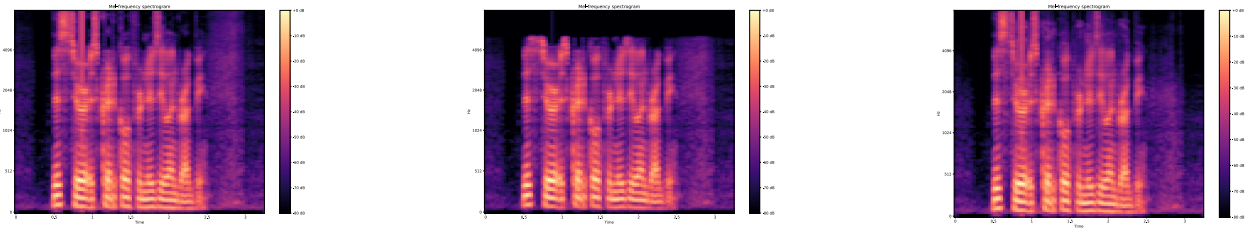


FIGURE 2. Black-box adversarial attack scenario. We introduce two augmentation techniques: audio denoising and band-pass filter. The CM model trained with augmented datasets is robust against adversarial samples generated by the surrogate model.

TABLE 1. Recent studies on adversarial attack and defense techniques in ASV-CM systems.

Study	Attack	Defense	Setting	ASV/CM System	Corpus	Year
[10]	FGSM, PGD	NA	Black/White-box	LCNN, SENet	ASVspoof 2019	2019
[11]	PGD	Spatial smoothing, Adversarial training	NA*	VGG, SENet	ASVspoof 2019	2020
[12]	FGSM, PGD	Mockingjay	Black-box	LCNN, SENet	ASVspoof 2019	2020
[13]	BIM, JSMA	VGG-like detection network	Black/White-box	GMM, TDNN, ResNet-34	Voxceleb1	2020
[14]	BIM	Self-supervised TERA	White-box (TERA aware/unaware)	ResNet, AAM-softmax	Voxceleb1, Voxceleb2	2021
[15]	APN, Voice Conversion	NA*	White-box	LCNN	ASVspoof 2019	2021
[16]	VC, Time-Domain Adversarial post-processing	NA*	White-box	ResNet-34	AIShell-3	2022
[17]	BIM	Vocoder for re-synthesis	NA*	NA*	Voxceleb1, Voxceleb2, Lrg	2022

(*) such information could not be precisely inferred from the studies



(a) original sample (b) 100-5k Hz sample (c) original sample after denoised

FIGURE 3. Mel-spectrograms of an audio waveform: original (a) , after high- and low-pass filtered (b); and after denoised (c).

clearer than the original waveform. We expect a good synthesizer to generate high-quality (speech) features in such a range, whereas the adversarial attack would exploit frequency ranges beyond human perception to maximize effectiveness. Logically, the use of narrower frequency ranges can defer the impact of the attack.

In order to find the suitable range of frequencies, we relied on the characteristics of the human auditory system. The human auditory system is most sensitive to the range between 100 Hz and 5 kHz. Whereas the frequency band around 2 kHz is the most crucial frequency range regarding perceived intelligibility. Therefore, we decided to mainly use the frequency range of 100-5k Hz for our experiments.[2]

[2]We also discuss other frequency ranges in the Evaluation section.

1) BAND-PASS FILTER PARAMETERS

In order to extract the preferred frequency ranges, low-pass and high-pass using *sinc* filters are used in conjunction with each other:

$$\text{sinc}(x) = \begin{cases} \frac{\sin(\pi x)}{\pi x}, & x \neq 0 \\ 1, & x = 0 \end{cases} \tag{3}$$

$$H_{LPF}(s) = \frac{\text{sinc}(W_c ts)}{\pi s}, \quad |s| \leq \pi W_c \tag{4}$$

$$H_{HPF}(s) = 1 - \frac{\text{sinc}(W_c ts)}{\pi s}, \quad |s| \leq \pi W_c \tag{5}$$

where W_c is the cutoff frequency, s is the complex frequency variable, and t is the sampling interval. Frequencies below the cutoff value of 5kHz are passed through the low-pass filter, whereas frequencies above 100Hz are passed through the high-pass filter.

C. MODEL ROBUSTNESS AGAINST ADVERSARIAL ATTACKS

Previous studies have used L_p clipping to control the quality of the attack audio sample [31]. For ASR systems, this approach may result in noisy audio [32]. CM systems, however, do not have such constraints. Since our assumption was that the attackers are able to manipulate the signal regarding psychoacoustics, we have made two important observations:

- CM models are responsible for spoofing detection, not for faithfully retaining the speech signal. Applying the band-pass filter just to capture the most significant signal features will not necessarily affect the CM's performance.
- Adversarial attacks can perturb high or low-frequency bands without making the audio perceived differently by humans. Approaches that utilize an entire frequency band may be vulnerable to such attacks.

Our study focuses on constructing robust CM models against adversarial samples by constraining the attack surface, specifically the perturbation capability of gradient-based attacks. Based on the investigation in Section III-B, we have identified 100-5k Hz frequency range is significant for general conversation. We anticipated that models trained with such a range will exhibit reduced vulnerability to adversarial samples.

It is important to note that, unlike previous studies, our approach addresses the perturbation space of the sample, rather than the degree of perturbation (ϵ) in each iteration. In other words, instead of controlling the audio quality, we solely govern *what a CM model can learn* by exclusively exposing it to a narrower band of frequencies, while disregarding frequencies beyond that range. We trained two types of neural networks on the augmented datasets: Squeeze-and-Excitation Networks (SENet) and Light Convolutional Neural Networks (LCNN) to see if adversarial robustness effect can be achieved. These models were later trained with denoised dataset for further evaluation. The feature used in this study is Log-Power spectrogram (LPS), extracted by Librosa [36]. The details are described in the next section.

D. AUDIO DENOISING

Perturbations made to the samples are assumed to be visually or audibly imperceptible to humans, and we consider them *adversarial noise* [37], [38]. There have been several studies in the computer vision domain where *denoising* is used to preprocess the image prior to feeding it into neural networks [37], [38]. In addition to the band-pass filter approach, we introduce denoising audio samples as an alternative augmentation method. The denoising is accomplished using VisuShrink, a soft thresholding technique that applies the universal threshold [39] to the wavelet coefficients, effectively separating the signal from the noise.

1) DENOISING PARAMETERS

For denoising, we used the following parameters.

- *block_size*: The size of each block in samples, set as 10% of the signal duration
- coefficients: The wavelet coefficients obtained by applying the discrete wavelet transform (DWT) on each block.
- *sigma*: The median absolute deviation of the wavelet coefficients
- *theshold*: The threshold used for the VisuShrink thresholding method.

The detailed configuration can be found at the source code which is available at our repository.[3]

E. EFFECTS OF MAX-FEATURE-MAPPING ON ADVERSARIAL ROBUSTNESS

Max-Feature-Map (MFM) operation [40] is introduced as a novel activation function for Convolutional Neural Networks (CNNs) to enhance feature extraction and robustness in the presence of noisy labels. By drawing inspiration from neural inhibition and maxout activation, MFM aims to achieve certain characteristics that can improve adversarial robustness and feature extraction in audio samples, even when the dataset contains noisy signals.

- Separation of Noisy and Informative Signals: In the context of audio samples, MFM can help separate noisy signals from informative signals. This means that during the training process, the neurons that receive noisy inputs are suppressed, while neurons that receive informative signals are activated, effectively enhancing the extraction of meaningful features from the audio data.
- Feature Extraction: Similar to lateral inhibition in neural science [41], MFM also achieves a form of inhibition to separate different types of information. For example, in the case of audio data, MFM emphasizes specific frequency components by suppressing the activity of neighboring frequency components, making it stand out more prominently in the audio signal.

[3]https://github.com/nguyenvulong/AdvDefenseCM

- Parameter-Free Inhibition: MFM does not depend on any learnable parameters during the training process. The inhibition process is fixed and does not involve any weights or biases that need to be updated through gradient-based optimization during training. This means that it does not rely heavily on training data, making it more resilient to variations and noise in the dataset. This lack of parameters in the inhibition process can contribute to adversarial robustness by avoiding overfitting to the training data and generalizing better to unseen samples.
- Compact and Light Models: MFM-based CNN models are designed to be light and efficient. By using the max function to suppress activations, MFM reduces the number of activated neurons and thus the overall complexity of the model.

In summary, these characteristics make MFM a promising activation function for improving the robustness and performance of CNNs on noisy audio data, which we empirically validated in the Evaluation section.

IV. EVALUATION

A. EXPERIMENT SETTINGS

1) MODELS

We used Light Convolutional Neural network (LCNN) [40] and Squeeze-and-Excitation Networks (SENet) [42] models,[4] as described in [10], to demonstrate their robustness against black-box adversarial attacks when trained with augmented techniques proposed in Section III. While squeeze-and-excitation block from SENet is known to improve the representational power of a network, LCNN is highly regarded in spoofing detection due to its compactness and efficacy. Later in this section, we also compared the robustness of these models with other recent studies like RawNet2 [28], AASIST-SSL [43] (a variant of ASSIST [44]), and BTS-E [45]. For better understanding of the impact of adversarial attacks and how well adversarial samples transfer between deep architectures, two variants of each networks were used.

- For LCNN, we used LCNN-large and LCNN-small, which have 10,198,816 and 509,072 hyperparameters, respectively.
- For SENet, we used SENet34 and SENet12, which contain 1,343,762 and 478,546 hyperparameters.

Similar to [10], this study also evaluated the robustness of these models in a black-box manner, where adversarial samples retrieved by attacking LCNN models are used to benchmark SENet models, and vice versa. We used notations **O**, **C**, and **D** to denote models trained with the **O**riginal, 100-5k Hz (band-pass filter with two **C**ut-off values), and **D**enoising datasets, respectively. For example, **SENet34 (C)** means that the SENet34 model has been trained with the frequency band of 100-5k Hz. Likewise, **LCNN (D)** indicates a model that was trained with denoised dataset. The success of a black-box attack depends on how such adversarial samples "transfer" from one neural network to another, which has been studied in the past [18], [19]. As the target models are spoofing detection systems, Equal Error Rate (EER%) was chosen over *accuracy*.

[4]Each model has two variants: large and small.

2) DATASETS

We used 5 well-known datasets to evaluate the proposed defense mechanisms. ASVSpoof 2019 and 2021 datasets that are mainly used to compare the robustness of band-pass filter against denoising approach in Tables 2, 3, 4 and 5. In-the-Wild (ItW) [46], Fake-or-Real [47], and ADD2023 [48] were used to further compare the performance of the augmented LCNN model with other state-of-the-art studies in Table 8.

For the training phase, as described in Fig. 2, we applied two augmentation techniques:

- Band-pass filter: We used *sox*[5] for waveform processing. Given two frequency cutoff values, *sinc* filter performs a low-pass filtering with higher value, followed by a high-pass filtering using lower value.
- Denoising: Thresholding technique[6] was employed to remove unwanted noisy signal.

For the evaluation phase, adversarial datasets were retrieved from FGSM and PGD attacks on different surrogate models. We categorize the datasets into several groups: Table 2 includes datasets that are generated by attacking several model instances of an LCNN, as described in IV-A1. For example, **FGSM LCNN small (O)** is an adversarial dataset generated by using FGSM to attack "LCNN small" model that was trained with the original dataset. Table 3 contains datasets generated by attacking SENet models. For instance, **FGSM SENet12 (C)** is an adversarial dataset generated by using FGSM to attack "SENet12" model that was trained with the band-pass frequencies with the two cutoff values at 100 and 5k Hz. Similarly, Table 4 and 5 refer to PGD attack and denoised datasets. We set perturbation degree ϵ to 5.0 in such attack scenarios.

B. MODEL ROBUSTNESS AGAINST BLACK-BOX ADVERSARIAL ATTACKS

In Tables 2, 3, 4 and 5, LCNN and SENet model instances were evaluated against different adversarial datasets. In the black-box setting, the surrogate models are expected to have different architectures from the target CM models. Adversarial examples generated by attacking the surrogate LCNN models are used to evaluate the SENet models and vice versa. In most cases, models trained with the original dataset have worse EERs than ones trained with the augmented datasets. The EER can be significantly reduced when the augmentation techniques are applied, e.g., in Table 4, the average EER (%) of the original model SENet34 (O) reduced from 39.68 to 17.39 and 14.02 when denoising and band-pass filter were

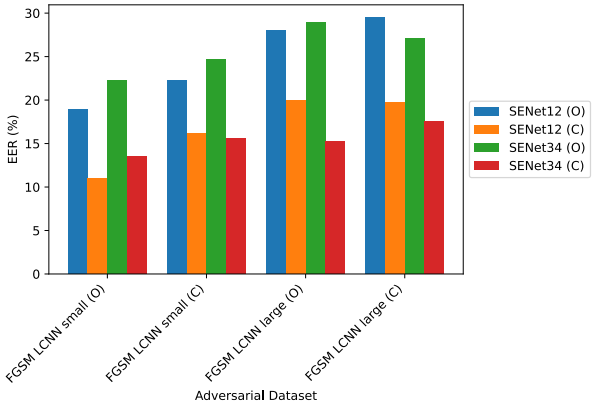
[5]https://sox.sourceforge.net
[6]https://github.com/AP-Atul/Audio-Denoising

TABLE 2. Compare between the original models and the models trained with 100-5k Hz frequency band using adversarial datasets generated from surrogate LCNN models. Target Models: SENet. Attack type: FGSM. Metric: EER(%).

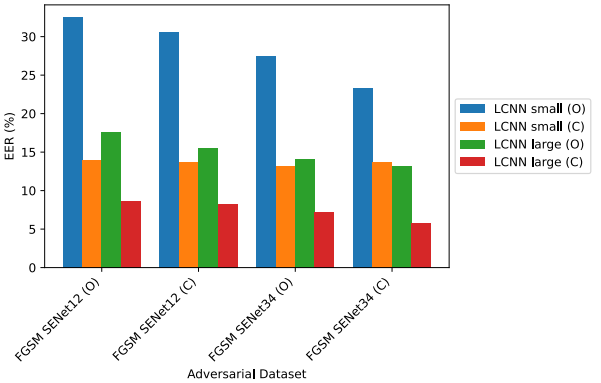
Dataset retrieved from adversarial attack	SENet12 (O)	SENet12 (C)	SENet34 (O)	SENet34 (C)
FGSM LCNN small (O)	18.87	**10.95**	22.22	13.48
FGSM LCNN small (C)	22.31	16.15	24.63	**15.53**
FGSM LCNN large (O)	27.97	19.99	28.98	**15.20**
FGSM LCNN large (C)	29.48	19.73	27.03	**17.51**
Average	**24.66**	**16.71**	**25.72**	**15.43**

TABLE 3. Compare between the original models and the models trained with 100-5k Hz frequency band using adversarial datasets generated from surrogate SENet models. Target Models: LCNN. Attack type: FGSM. Metric: EER(%).

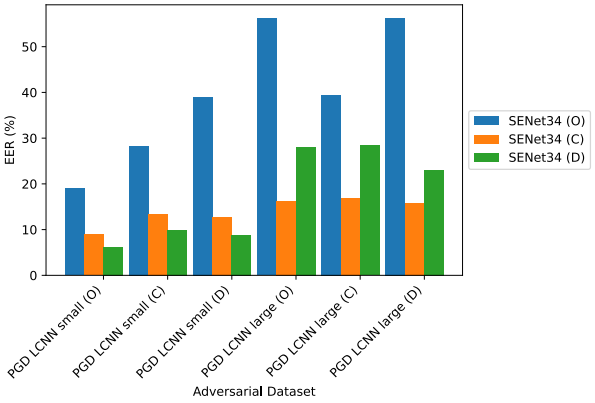
Dataset retrieved from adversarial attack	LCNN small (O)	LCNN small (C)	LCNN large (O)	LCNN large (C)
FGSM SENet12 (O)	32.50	14.01	17.55	**8.62**
FGSM SENet12 (C)	30.65	13.64	15.49	**8.25**
FGSM SENet34 (O)	27.51	13.13	14.05	**7.19**
FGSM SENet34 (C)	23.32	13.67	13.13	**5.75**
Average	**28.50**	**13.61**	**15.06**	**7.45**



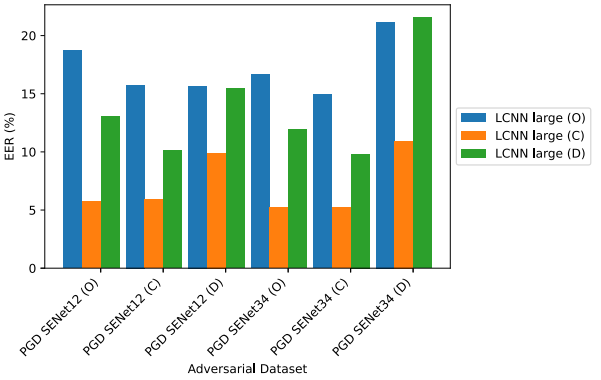
(a) FGSM attack on SENet models



(b) FGSM attack on LCNN models



(c) PGD attack on SENet models



(d) PGD attack on LCNN models

FIGURE 4. Black-box Attacks on SENet and LCNN models: (O), (C), and (D) denote models trained with the original frequency band, 100Hz - 5kHz band, and denoised datasets, respectively. The lower the EER(%), the better.

applied. Similarly, in Table 5, the EER (%) of the LCNN (O) dropped from 17.15 to 13.66 and 7.33 when trained with the augmented datasets.

In Fig. 4, the LCNN models trained with frequency band 100-5kHz performed consistently better than those trained with the denoised dataset. In particular, the LCNN model (C) had an average EER (%) of 7.33, compared to 13.66 in the case of the denoising approach (LCNN (D)). The Max-Feature-Map (MFM) in LCNN models [40] contributes to robustness against noisy samples without the requirement of denoising, resulting in LCNN large (C) achieving the highest overall performance with band-pass augmentation.

TABLE 4. Comparison between original, denoising and band-pass filter approaches. Surrogate models: LCNN. Target models: SENet. Attack type: PGD. Metric: EER(%).

Dataset retrieved from adversarial attack	SENet34 (O)	SENet34 (C)	SENet34 (D)
PGD LCNN small (O)	18.99	8.95	6.19
PGD LCNN small (C)	28.30	13.37	9.85
PGD LCNN small (D)	38.93	12.83	8.83
PGD LCNN large (O)	56.13	16.22	28.01
PGD LCNN large (C)	39.41	16.94	28.39
PGD LCNN large (D)	56.34	15.80	23.08
Average	**39.68**	**14.02**	**17.39**

TABLE 5. Comparison between original, denoising and band-pass filter approaches. Surrogate models: SENet. Target models: LCNN. Attack type: PGD. Metric: EER(%).

Dataset retrieved from adversarial attack	LCNN large (O)	LCNN large (C)	LCNN large (D)
PGD SENet12 (O)	18.72	5.71	13.07
PGD SENet12 (C)	15.75	5.88	10.14
PGD SENet12 (D)	15.68	9.92	15.45
PGD SENet34 (O)	16.68	5.22	11.91
PGD SENet34 (C)	14.94	5.23	9.83
PGD SENet34 (D)	21.18	10.94	21.57
Average	**17.15**	**7.33**	**13.66**

During our analysis of the Signal-to-Noise Ratio (SNR), the adversarial signal's power seemed to change significantly compared to the original signal's. This result is understandable because this attack aims to deceive the classifier (i.e., countermeasure system). In the case of band-passed and denoised signals, their signal powers seemed similar but actually differed when considering the noise removed. Band-passed signal has a negative value of SNR, which can be explained by the thin band of frequencies (100Hz-5kHz) representing the signal, whereas the noise was taken from all frequencies. Also, the SNR between the denoised and adversarial signals is different. In the denoising technique, only noisy signals are reduced or removed. While this technique can somewhat alleviate the adversarial noise added by the attacker, it cannot completely remove the adversarial noise since the whole waveform can still be perturbed. On the other hand, the proposed band-pass filter proactively prohibits all perturbations on high frequencies, effectively forcing the neural networks to learn discriminative features in the proposed range of 100Hz-5kHz. This approach has a much stronger effect on noise reduction.

For a relative comparison, adversarial robustness can be evaluated by different metrics. The authors of [11] and [12] used *accuracy*. After adversarial training against the PGD attack, the accuracy of SENet improved from 78.93% to 83.72% (Gaussian filter) [11]. Also, [13] measured detection accuracy in case of active defense against BIM attack. For instance, the detection accuracy was 48.61% for $\epsilon = 0.3$. The authors of [14] and [17] managed to reduce the EER by 43%, approximately.

Under the same settings[7] (ϵ = 5.0, PGD SENet12 (O) dataset), we compared our defense against the PGD attack from [10]; the LCNN models in Table 5 reduced the EER from 18.72% to 5.71% and 13.07%, respectively. The similar impact of the augmentation techniques on other model instances can also be inferred from the two Tables 4 and 5.

[7]See Table 4 (b) in [10].

C. COMPARISON WITH OTHER MODELS

This section evaluates the deception rates of adversarial samples against recent models proposed in the literature. We chose a LCNN model trained with the proposed frequency band-pass as a candidate. There were 40,959 audio samples from ASVSpoof 2021 [49] that have been used for launching PGD attack on different models in Table 6. We first started with white-box attack, similar to what have been done in previous experiments, to retrieve adversarial candidates. We selected samples that were previously and correctly classified as *spoof* by a model, and got misclassified by the same model after the white-box attack. These samples were then used to attack other models in black-box manner.

As described in Table 6, AASIST-SSL showed a very strong resistance to adversarial samples except for the white-box case where the deception rate increased to 26.71%, which is understandable in deep learning context. The proposed PGD LCNN large (C) model was robust against all adversarial datasets, including the white-box scenario, where the deception rate was only 0.08%. For other cases, RawNet2 seemed to be vulnerable to the attacks from other models except the LCNN; BTS-E was also built on top of RawNet2 but adopted several biological components such as breathing and silence and it does not suffer from the same limitation. We also observed that BTS-E and LCNN were incapable of generating high quality adversarial examples (i.e., having low deception rates with all models). Such phenomenon might have been related to their robustness. LCNN leveraged Max-Feature-Map to remove noisy data, whereas BTS-E fused several feature types for detection.

TABLE 6. Comparison between the robustness of recent models. Metric: Deception Rate (%). Attack type: PGD. Adversarial samples were generated by attacking the models and then reused in black-box manner.

	PGD_LCNN large (C)	PGD_RawNet2	PGD_BTS-E	PGD_ASSIST-SSL
LCNN_large (C)	0.08	5.05	3.16	2.92
RawNet2 [28]	3.57	29.78	26.09	18.02
BTS-E [45]	3.95	6.30	2.50	4.08
AASIST-SSL [43] [44]	0.08	3.10	8.37	26.71

To provide a better view of the quality of the adversarial samples, we presented the predicted MOS (Mean Opinion Score) of the generated samples in Table 7. While *high* speech quality does not mean *better* deception rate, perceptibly distorted audio samples would decrease MOS scores. The LCNN model did not generate a lot of deceptive samples, which is why the distribution of MOS scores was very close to the original samples. It is important to note that the MOS of *original samples* from the ASVSpoof dataset (about 60% in our experiment) is below 3. Therefore we could not expect the adversarial samples to have higher speech quality (i.e., high MOS scores).

In the final analysis, Table 8 presents the results of generating adversarial samples from fake speech datasets, specifically In-the-Wild [46], Fake-or-Real [47], and ADD2023 [48], through white-box attacks on the LCNN model. We utilized these adversarial samples to evaluate the efficacy of our band-pass filter defense, alongside other state-of-the-art studies. Notably, the LCNN model exhibits significant resilience against direct attacks. Two datasets FoR and ADD2023 achieved 0% deception rate, i.e., no successful adversarial samples were found. For the in-the-wild adversarial dataset, all four models still showed their resistance against the attack besides some successful cases of deception. Even though the selected datasets were small in size, it is crucial to note that in this context, we employed a black-box benchmarking approach for other models such as RawNet2, BTS-E, and AASIST-SSL, while the evaluation of the LCNN model was carried out using a white-box manner.

D. DISCUSSION

1) DIFFERENT FREQUENCY BANDS

It is logical to argue that using smaller frequency bands would yield similar effects since the adversarial noise introduced by perturbation can be further restricted. In this experiment, we retrained the models from scratch on a dataset with narrower frequency bands (i.e., 2-4kHz). The performance of models trained on these smaller ranges was better than when using the full range, but it did not surpass the performance achieved with the proposed range (i.e., 100Hz - 5kHz). For instance, the best performance (EER%) of a model using the proposed range is 7.00, whereas the model operating within the 2-4kHz frequency range is 13.81 under the same dataset. Our understanding suggests that frequency ranges smaller than the proposed range possess fewer discriminative features, aligning with Section I where it is indicated that the auditory system exhibits heightened sensitivity starting at 500Hz. Smaller frequency ranges such as 2-4kHz failed to capture such details. We also verified that a model learns less information beyond the range 100-5k Hz. For instance, the best model trained with frequencies beyond 100-5k Hz scored 39.55% on ASVspoof 2021 dataset [49], whereas the one trained with the proposed frequency band scored 21.68%.

Besides, the prominence of specific frequency bands can vary based on factors such as gender and speech contents. It is important to note that our objective was not to precisely capture all speech contents from different genders, a critical aspect in tasks like speaker verification, but of lesser importance in the context of spoofing detection. We acknowledge the potential benefits of utilizing dynamic frequency bands tailored to specific characteristics like gender or age groups. However, we chose to utilize a unified frequency range for several reasons: (1) By using a unified range of frequency, we reduce the complexity and computational overhead associated with training multiple frequency bands and classifying samples into different groups based on gender or age; (2) Our current research serves as a foundation for future work, where we aim to explore and extend the proposed approach to address diverse languages and dialects. Having a simple yet unified frequency range allows us to focus on the broader problem of spoofing detection and ensures our method's applicability to various scenarios.

2) OTHER ADVERSARIAL ATTACKS

During experiments, we were aware of Carlini and Wagner attack [5], also known as C&W, which is a very strong type of adversarial attacks. The authors have successfully launched the targeted attack on Speech-to-Text systems. However, we realized that using C&W for generating adversarial samples was time-consuming (100-120 minutes per sample on a Quadro RTX 5000). Therefore, we did not investigate further the effect of this attack.

We also evaluated the models against Momentum Iterative Fast Gradient Sign Method (MI-FGSM) [50] and found that AASIST performance can be worsen. The experiment was implemented using *torchattacks* [51]. As reported in Table 9, the augmented LCNN model was also robust against MI-FGSM attack, whereas RawNet2 and AASIST-SSL suffered from a significant degradation with 22.99% and 19.67% deception rates. Interestingly, BTS-E was not affected much by this kind of attack.

3) MODEL GENERALIZATION

In the ASVSpoof 2021 challenge,[8] the EER(%) scores of the DF track (*eval* subset) ranged from 15.64 to 29.75,

[8]Deepfake Track.

TABLE 7. MOS score distribution (Unit %) of generated adversarial examples. MOS score ≥ 4 indicates a very good quality of speech.

	MOS ≤ 1	1 <MOS ≤ 2	2 <MOS ≤ 3	3 <MOS ≤ 4	MOS >4
PGD_LCNN large (C)	0.00	16.30	36.67	33.85	13.16
PGD_RawNet2	21.83	69.78	8.35	0.03	0.00
PGD_BTS-E	19.40	75.43	5.16	0.00	0.00
PGD_ASSIST-SSL	6.89	88.97	4.13	0.00	0.00
Original samples	1.05	24.87	39.15	26.27	8.64

TABLE 8. Comparison between the robustness of recent models. Metric: Deception Rate (%). Attack type: PGD. Adversarial samples from different datasets were re-generated by attacking the LCNN model itself.

	ItW [46]	**FoR** [47]	**ADD2023** [48]
LCNN (this study)	0.81	0	0.5
RawNet2 [28]	0.41	0.9	0
BTSE [45]	1.21	2.25	0
AASIST-SSL [43] [44]	0.41	0	0

TABLE 9. MI-FGSM attack on SENet34 model. The retrieved adversarial samples were used for this evaluation.

	LCNN (this study)	RawNet2	BTS-E	AASIST-SSL
Deception Rate (%)	0	22.99	0.15	19.67

TABLE 10. Comparison with state-of-the-art studies. The models were evaluated against original synthetic speech (without adversarial attacks).

	LCNN (this study)	RawNet2	AASIST-SSL	BTS-E
Accuracy	95.48%	91.37%	98.69%	94.01%

where as the one trained with the proposed frequency band scored 21.68.

Another test that we conducted was evaluating the models against original synthetic speech of ASVSpoof 2021, that is, we did not introduce adversarial samples to this experience. The result reported in Table 10 showed that the augmented LCNN can still perform well on normal dataset, right behind the AASIST-SSL.

4) COMPUTATIONAL COMPLEXITY ANALYSIS

$$O(E \times \frac{N}{B} \times (P_f + P_b + L)) \tag{6}$$

where E is the number of epochs; N is the total number of samples, up to 70,000 samples per adversarial dataset; B is the batch size (from 64 to 512 in our experiment); P_f and P_b is the complexity of the model's forward and backward passes (proportionate to the number of hyper-parameters), respectively; L is the complexity of the loss function (per batch), for example, A-softmax [52].

5) LIMITATIONS

Our primary focus was on developing countermeasures (CM) that can adeptly tackle the spoofing detection task, without delving into the intricate details required for tasks such as precise speaker verification across different genders. To achieve this, we opted for a frequency range that encapsulates significant speaker information while avoiding unnecessary computational complexities. However, the potential advantages of employing dynamic frequency bands tailored to specific characteristics, such as gender or age groups, were not extensively explored, which represents a limitation in our study.

Also, as discussed in the previous section, there is a need to enhance the model's performance beyond the adversarial scenario to ensure effective generalization to in-the-wild datasets, which was not fully explored in this study.

Furthermore, it is important to acknowledge that our investigation primarily concentrated on black-box adversarial transferred attacks, and we could further enhance our work by giving due attention to the white-box scenario. In conclusion, while our proposed approach effectively addresses the spoofing detection task, we recognize the need for further exploration and expansion of our methods to encompass a wider range of scenarios.

V. CONCLUSION

In this study, we explored frequency band-pass filter and denoising techniques, the two potential defenses against black-box adversarial attacks on spoofing countermeasure models. The first scheme offered promising results on several adversarial datasets, whereas the latter still has room for improvement. We empirically showed that band-pass filter is a simple yet effective augmentation technique to enhance the security of CM systems. We also note that the white-box setting and model generalization were not fully explored in this work, which represents a limitation of our research that we aim to address in future studies.

REFERENCES

[1] M. Sahidullah, H. Delgado, M. Todisco, A. Nautsch, X. Wang, T. Kinnunen, N. Evans, J. Yamagishi, and K.-A. Lee, "Introduction to voice presentation attack detection and recent advances," in *Handbook of Biometric Anti-Spoofing: Presentation Attack Detection and Vulnerability Assessment*. Berlin, Germany: Springer, 2023, pp. 339–385.

[2] X. Wang and J. Yamagishi, "A practical guide to logical access voice presentation attack detection," in *Frontiers in Fake Media Generation and Detection*. Berlin, Germany: Springer, 2022, pp. 169–214.

[3] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, "Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward," *Appl. Intell.*, vol. 53, no. 4, pp. 3974–4026, 2022.

[4] M. M. Cisse, Y. Adi, N. Neverova, and J. Keshet, "Houdini: Fooling deep structured visual and speech recognition models with adversarial examples," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[5] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *Proc. IEEE Secur. Privacy Workshops (SPW)*, May 2018, pp. 1–7.

[6] L. Schonherr, K. Kohls, S. Zeiler, T. Holz, and D. Kolossa, "Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, San Diego, CA, USA, 2019, pp. 1–18.

[7] Y. Qin, N. Carlini, G. Cottrell, I. Goodfellow, and C. Raffel, "Imperceptible, robust, and targeted adversarial examples for automatic speech recognition," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 5231–5240.

[8] P. Neekhara, S. Hussain, P. Pandey, S. Dubnov, J. McAuley, and F. Koushanfar, "Universal adversarial perturbations for speech recognition systems," in *Proc. Interspeech*, Sep. 2019.

[9] Z. Li, C. Shi, Y. Xie, J. Liu, B. Yuan, and Y. Chen, "Practical adversarial attacks against speaker recognition systems," in *Proc. 21st Int. Workshop Mobile Comput. Syst. Appl.*, Mar. 2020, pp. 9–14.

[10] S. Liu, H. Wu, H.-Y. Lee, and H. Meng, "Adversarial attacks on spoofing countermeasures of automatic speaker verification," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2019, pp. 312–319.

[11] H. Wu, S. Liu, H. Meng, and H.-Y. Lee, "Defense against adversarial attacks on spoofing countermeasures of ASV," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6564–6568.

[12] H. Wu, A. T. Liu, and H.-Y. Lee, "Defense for black-box attacks on anti-spoofing models by self-supervised learning," in *Proc. Interspeech*, Oct. 2020, pp. 3780–3784.

[13] X. Li, N. Li, J. Zhong, X. Wu, X. Liu, D. Su, D. Yu, and H. Meng, "Investigating robustness of adversarial samples detection for automatic speaker verification," in *Proc. Interspeech*, Oct. 2020, pp. 1–5.

[14] H. Wu, X. Li, A. T. Liu, Z. Wu, H. Meng, and H.-Y. Lee, "Adversarial defense for automatic speaker verification by cascaded self-supervised learning models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6718–6722.

[15] Y.-Y. Ding, L.-J. Liu, Y. Hu, and Z.-H. Ling, "Adversarial voice conversion against neural spoofing detectors," in *Proc. Interspeech*, Aug. 2021, pp. 816–820.

[16] C. Wen, T. Guo, X. Tan, R. Yan, S. Zhou, C. Xie, W. Zou, and X. Li, "Time domain adversarial voice conversion for ADD 2022," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 9221–9225.

[17] H. Wu, P.-C. Hsu, J. Gao, S. Zhang, S. Huang, J. Kang, Z. Wu, H. Meng, and H.-Y. Lee, "Adversarial sample detection for speaker verification by neural vocoders," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 236–240.

[18] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proc. ACM Asia Conf. Comput. Commun. Secur.*, Apr. 2017, pp. 506–519.

[19] K. Michel Koerich, M. Esmailpour, S. Abdoli, A. D. S. Britto, and A. L. Koerich, "Cross-representation transferability of adversarial attacks: From spectrograms to audio waveforms," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–7.

[20] S. Hussain, P. Neekhara, S. Dubnov, J. McAuley, and F. Koushanfar, "WaveGuard: Understanding and mitigating audio adversarial examples," in *Proc. 30th USENIX Secur. Symp.*, 2021, pp. 2273–2290.

[21] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. 3rd Int. Conf. Learn. Represent.*, Y. Bengio and Y. LeCun, Eds. San Diego, CA, USA, 2015, pp. 1–11.

[22] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. 6th Int. Conf. Learn. Represent.*, Vancouver, BC, Canada, 2018, pp. 1–28.

[23] R. K. Das, X. Tian, T. Kinnunen, and H. Li, "The attacker's perspective on automatic speaker verification: An overview," in *Proc. Interspeech*, Oct. 2020, pp. 1–5.

[24] J. Villalba, Y. Zhang, and N. Dehak, "X-vectors meet adversarial attacks: Benchmarking adversarial robustness in speaker verification," in *Proc. Interspeech*, Oct. 2020, pp. 4233–4237.

[25] Y. Zhang, Z. Jiang, J. Villalba, and N. Dehak, "Black-box attacks on spoofing countermeasures using transferability of adversarial examples," in *Proc. Interspeech*, Oct. 2020, pp. 4238–4242.

[26] A. Kassis and U. Hengartner, "Practical attacks on voice spoofing countermeasures," 2021, *arXiv:2107.14642*.

[27] T. Chen, A. Kumar, P. Nagarsheth, G. Sivaraman, and E. Khoury, "Generalization of audio deepfake detection," in *Proc. Speaker Lang. Recognit. Workshop (Odyssey)*, Nov. 2020, pp. 132–137.

[28] H. Tak, J.-W. Jung, J. Patino, M. Kamble, M. Todisco, and N. Evans, "End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection," in *Proc. Autom. Speaker Verification Spoofing Countermeasures Challenge*, 2021, pp. 1–8.

[29] W. Ge, J. Patino, M. Todisco, and N. Evans, "Raw differentiable architecture search for speech deepfake and spoofing detection," in *Proc. Ed. Autom. Speaker Verification Spoofing Countermeasures Challenge*, Sep. 2021, pp. 22–28.

[30] I.-Y. Kwak, S. Choi, J. Yang, Y. Lee, S. Han, and S. Oh, "Low-quality fake audio detection through frequency feature masking," in *Proc. 1st Int. Workshop Deepfake Detection Audio Multimedia*, Oct. 2022, pp. 9–17.

[31] H. Abdullah, K. Warren, V. Bindschaedler, N. Papernot, and P. Traynor, "SoK: The faults in our ASRs: An overview of attacks against automatic speech recognition and speaker identification systems," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2021, pp. 730–747.

[32] H. Abdullah, M. S. Rahman, C. Peeters, C. Gibson, W. Garcia, V. Bindschaedler, T. Shrimpton, and P. Traynor, "Beyond L_p clipping: Equalization based psychoacoustic attacks against ASRs," in *Proc. Asian Conf. Mach. Learn.*, 2021, pp. 672–688.

[33] B. Liu, J. Zhang, and J. Zhu, "Boosting 3D adversarial attacks with attacking on frequency," *IEEE Access*, vol. 10, pp. 50974–50984, 2022.

[34] M. J. Antuñano and J. P. Spanyers, "Hearing and noise in aviation," FAA Civil Aerospace Med. Inst. Aeromedical Educ. Division, OK, USA, Tech. Rep. AM-400-98/3, 2006. [Online]. Available: https://www.faa.gov/pilots/safety/pilotsafetybrochures/media/hearing.pdf

[35] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee, and L. Juvela, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Comput. Speech Lang.*, vol. 64, Nov. 2020, Art. no. 101114.

[36] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "Librosa: Audio and music signal analysis in Python," in *Proc. 14th Python Sci. Conf.*, 2015, pp. 18–25.

[37] C. Xie, Y. Wu, L. v. d. Maaten, A. L. Yuille, and K. He, "Feature denoising for improving adversarial robustness," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 501–509.

[38] Y. Bakhti, S. A. Fezza, W. Hamidouche, and O. Déforges, "DDSA: A defense against adversarial attacks using deep denoising sparse autoencoder," *IEEE Access*, vol. 7, pp. 160397–160407, 2019.

[39] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, Sep. 1994.

[40] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 11, pp. 2884–2896, Nov. 2018.

[41] E. Fiesler and R. Beale, *Handbook of Neural Computation*. Boca Raton, FL, USA: CRC Press, 2020.

[42] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[43] H. Tak, M. Todisco, X. Wang, J.-W. Jung, J. Yamagishi, and N. Evans, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," 2022, *arXiv:2202.12233*.

[44] J.-W. Jung, H.-S. Heo, H. Tak, H.-J. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, "AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 6367–6371.

[45] T.-P. Doan, L. Nguyen-Vu, S. Jung, and K. Hong, "BTS-E: Audio deepfake detection using breathing-talking-silence encoder," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[46] N. Müller, P. Czempin, F. Diekmann, A. Froghyar, and K. Böttinger, "Does audio deepfake detection generalize?" in *Proc. Interspeech*, Sep. 2022, pp. 1–5.

[47] R. Reimao and V. Tzerpos, "FoR: A dataset for synthetic speech detection," in *Proc. Int. Conf. Speech Technol. Human-Comput. Dialogue (SpeD)*, Oct. 2019, pp. 1–10.

[48] J. Yi, J. Tao, R. Fu, X. Yan, C. Wang, T. Wang, C. Y. Zhang, X. Zhang, Y. Zhao, Y. Ren, and L. Xu, "ADD 2023: The second audio deepfake detection challenge," 2023, *arXiv:2305.13774*.

[49] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch, and K. A. Lee, "ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 2507–2522, 2023.

[50] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9185–9193.
[51] H. Kim, "Torchattacks: A PyTorch repository for adversarial attacks," 2020, *arXiv:2010.01950*.
[52] Y. Li, F. Gao, Z. Ou, and J. Sun, "Angular softmax loss for end-to-end speaker verification," in *Proc. 11th Int. Symp. Chin. Spoken Lang. Process. (ISCSLP)*, Nov. 2018, pp. 190–194.

LONG NGUYEN-VU (Member, IEEE) received the B.S. degree in computer science from the National University of Information Technology, Ho Chi Minh, Vietnam, and the M.S. and Ph.D. degrees in engineering from Soongsil University, Seoul, South Korea. Before joining the Communication Network Security Laboratory, Soongsil University, he was a System Engineer with VNG, Vietnam, a game company. He is currently a Postdoctoral Researcher with Soongsil University. His current research interests include information security, applied machine learning, and cloud computing.

THIEN-PHUC DOAN (Student Member, IEEE) received the joint B.S. degree in cyber security from Vietnam National University, Ho Chi Minh City, and the University of Information Technology, in 2018, and the M.E. degree in electronic engineering from Soongsil University, in 2020. His current research interests include web security, android malware analysis, and deepfake speech synthesis and detection.

MAI BUI received the B.S. degree in information technology from the National University of Information Technology, Ho Chi Minh, Vietnam, in 2021. She is currently pursuing the master's degree with Soongsil University, Seoul, South Korea. Her current research interests include machine learning applications and deepfake speech synthesis and detection.

KIHUN HONG (Member, IEEE) received the B.S. degree in electronics engineering and the M.S. and Ph.D. degrees in information and telecommunication engineering from Soongsil University, Seoul, South Korea, in 2000, 2002, and 2006, respectively. From 2006 to 2007, he was a Postdoctoral Researcher with the University of California at Davis, USA. He was a Principal Engineer with Secui, South Korea, from 2008 to 2022, a network security company. He is currently an Assistant Professor with the School of Electronic Engineering, Soongsil University. His current research interests include deepfake detection, zero trust architecture, and industrial cybersecurity.

SOUHWAN JUNG (Member, IEEE) received the B.S. and M.S. degrees in electronics engineering from Seoul National University, in 1985 and 1987, respectively, and the Ph.D. degree from the University of Washington, Seattle, USA, in 1996. From 1996 to 1997, he was a Senior Software Engineer with Stellar One Corporation, Bellevue, USA. In 1997, he joined the School of Electronic Engineering, Soongsil University, Seoul, South Korea, where he is currently a Professor. He was the Research and Development Program Director of the Ministry of Knowledge Economy, South Korea, from 2009 to 2011, for information security area. He is also an Executive Director of the Korea Institute of Information Security and Cryptology. Since 2012, he has been leading the Smart Security Service Research Center funded by Korea Government, for six years. His current research interests include machine learning security, cloud security, and the IoT security.