IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 19 December 2022, accepted 30 December 2022, date of publication 4 January 2023, date of current version 9 January 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3234110

RESEARCH ARTICLE

Exploring Diverse Feature Extractions for Adversarial Audio Detection

YUJIN CHOI, JINSEONG PARK, JAEWOOK LEE, AND HOKI KIM
Department of Industrial Engineering, Seoul National University, Seoul 08826, South Korea

Corresponding author: Hoki Kim (ghrl9613@snu.ac.kr)

This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) Grant funded by the Korean Government (Ministry of Science and Information & Communications Technology, MSIT) (No.2022-0-00984, Development of Artificial Intelligence Technology for Personalized Plug-and-Play Explanation and Verification of Explanation), the National Research Foundation of Korea (NRF) Grant funded by the Korean Government (Ministry of Science and Information & Communications Technology, MSIT) (No. 2019R1A2C2002358), and the National Research Foundation of Korea (NRF) Grant funded by the Korean Government (Ministry of Science and Information & Communications Technology, MSIT) (No. 2022R1A5A6000840).

ABSTRACT Although deep learning models have exhibited excellent performance in various domains, recent studies have discovered that they are highly vulnerable to adversarial attacks. In the audio domain, malicious audio examples generated by adversarial attacks can cause significant performance degradation and system malfunctions, resulting in security and safety concerns. However, compared to recent developments in the audio domain, the properties of the adversarial audio examples and defenses against them still remain largely unexplored. In this study, to provide a deeper understanding of the adversarial robustness in the audio domain, we first investigate traditional and recent feature extractions in terms of adversarial attacks. We show that adversarial audio examples generated from different feature extractions exhibit different noise patterns, and thus can be distinguished by a simple classifier. Based on the observation, we extend existing adversarial detection methods by proposing a new detection method that detects adversarial audio examples using an ensemble of diverse feature extractions. By combining the frequency and self-supervised feature representations, the proposed method provides a high detection rate against both white-box and black-box adversarial attacks. Our empirical results demonstrate the effectiveness of the proposed method in speech command classification and speaker recognition.

INDEX TERMS Adversarial robustness, speech classification, feature extraction, adversarial example detection.

I. INTRODUCTION

Recent advances in deep learning have demonstrated significant performance improvements in various domains such as computer vision and speech recognition, yielding a large number of industrial applications [1], [2]. In particular, by combining deep learning models and traditional signal processing techniques (e.g., Mel-spectrogram), speech recognition systems have been successfully developed to identify or verify the physical sound of a human voice [3], [4]. More recently, a line of work on self-supervised learning method (e.g., Wav2vec [5]) has also improved the performance on speech recognition tasks by converting an original waveform to a feature with only neural network-based models. Based on these improvements, deep learning models are now actively used in real-world applications such as autonomous vehicle and smart home devices.

However, recent studies have revealed that deep learning models are vulnerable to adversarial attacks that generate malicious examples with subtle noises [6], [7]. The potential risks of deep learning models can be induced by adversarial attacks in real-world applications [8], [9], and the audio domain is no exception. Indeed, recent studies have verified that speech classification systems such as speech command classification and speaker recognition models can easily malfunction due to attacks by adversarial audio examples [10], [11], [12]. Figure 1 illustrates how adversarial attacks generate malicious audio examples on speech classification models. Given an input waveform $\boldsymbol{x}$ and the corresponding label y (or a one-hot vector $\boldsymbol{y}$), adversarial

The associate editor coordinating the review of this manuscript and approving it for publication was Ghulam Muhammad.

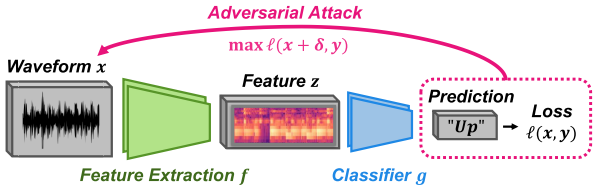
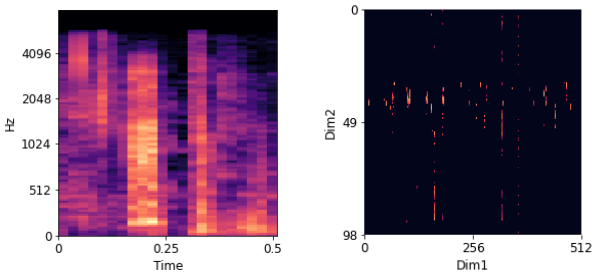


FIGURE 1. Illustration of adversarial attacks on speech classification tasks.



(a) Mel-spectrogram (b) Wav2vec

FIGURE 2. Visualization of Mel-spectrogram and Wav2vec features. Both features are extracted from the same waveform x, but have different feature dimensions and sparsity.

attacks aim to maximize the final loss $\ell(\boldsymbol{x}, \boldsymbol{y})$ by adding noise $\boldsymbol{\delta}$ to the waveform $\boldsymbol{x}$. Since the attack process can be easily done by using propagation, these types of adversarial attacks in the audio domain could entail severe hazards, e.g., causing autonomous vehicle accidents by manipulating voice commands or extracting private information by circumventing voice authorization.

There have been many studies conducted to develop adversarial attacks on speech recognition tasks in order to understand adversarial robustness in the audio domain. Following the Carlini and Wagner attack [11], various attacks have revealed the vulnerability of speech recognition models [10], [13], [14]. To defend models against adversarial audio examples, numerous researchers have proposed defense methods [15], [16]. However, most adversarial attacks and defenses have mainly been evaluated on models with traditional feature extractions, such as Mel-spectrogram. The analysis of models with newly proposed feature extractions, such as Wav2vec, is necessary to realize robustness guarantees in the audio domain.

In this work, to gain a deeper understanding of adversarial robustness in the audio domain, we analyze adversarial robustness of speech recognition models with traditional and newly proposed feature extractions. We first compare the adversarial robustness of models using different feature extractions, and investigate generated adversarial audio examples. We find that there exist different noise patterns of adversarial audio examples generated from Mel-spectrogram and Wav2vec, and demonstrate that they are easily distinguished by a simple classifier due to their difference. Lastly, based on the observation, we propose a new detection method that distinguishes adversarial audio examples from benign examples. By using an ensemble of models comprising traditional frequency extraction and modern latent representation extractions, and comparing the distance of outputs between models, the proposed method successfully detects adversarial audio examples compared to previous methods. The experimental results confirm that the proposed method achieves the state-of-the-art detection performance on speech command classification and speaker recognition tasks.

II. RELATED WORKS

A. SPEECH CLASSIFICATION

Speech classification systems transform an audio waveform $\boldsymbol{x}$ to a correct label y (or a one-hot vector $\boldsymbol{y}$). To utilize an audio waveform as input, a speech classification system has a *feature extraction* and *classifier* (illustrated in Figure 1). Feature extractions generally map an input waveform to a numeric representation vector, in order to extract important characteristics. Given a feature extraction f, a feature $\boldsymbol{z} = f(\boldsymbol{x})$ is extracted from the original waveform $\boldsymbol{x}$, and fed to a classifier g. Then, g outputs the probability vector $\boldsymbol{p} = g(\boldsymbol{z})$. Generally, in classification tasks, the i-th component that has the maximum value, i.e. $\arg\max_i p_i$, is used as a prediction $\hat{y}$. As the classifier g, deep learning classifiers are considered powerful baselines in speech classification.

Traditionally, frequency extraction methods such as Mel-spectrogram and Mel-frequency cepstral coefficient have been widely used [3], [4]. Frequency-based methods usually involve dividing the waveform into several overlapping frames, then transforming each frame into a spectrogram. To do this, a fast Fourier transform (FFT) is applied to the frame, followed by windowing and Mel-filtering. This produces a Mel-spectrogram feature, from which the Mel-frequency cepstral coefficients (MFCC) are extracted using the discrete cosine transform (DCT). In Figure 2, we show an example of a Mel-spectrogram feature. These types of transforms allow us to handle the original waveform by extracting useful features and achieve human-level performance in various speech classification tasks [4], [17], [18], [19].

Recently, Schneider et al. [5] proposed a self-supervised model called Wav2vec. Wav2vec is a set of convolutional neural networks trained on a large amount of raw audio data without any frequency extractions. In short, Wav2vec aims to distinguish a feature vector (or a latent representation) $\boldsymbol{z}$ from other feature vectors $\tilde{\boldsymbol{z}}$ extracted from different audios by using contrastive losses. Finally, a set of front layers (i.e., the encoder) of Wav2vec is used as a feature extraction f. Unless otherwise noted, Wav2vec implies the encoder in the rest of the paper. We plot an example of extracted feature vector $\boldsymbol{z}$ from Wav2vec in Figure 2b. Wav2vec produces a more sparse and higher dimensional feature vector, which significantly differs from the Mel-spectrogram (Figure 2a). Based on the advantages of self-supervised training, Wav2vec trains numerous audio samples and harness an

effective latent space that largely improves the performance on variety speech recognition tasks. Following the success of Wav2vec, many researchers have proposed task-specific Wav2vec-based models [20].

Although recent studies on speech recognition systems have used Wav2vec-based models, only a few works have addressed the adversarial robustness of these models [21]. In this work, we analyze adversarial robustness of Wav2vec-based models and compare it to Mel-spectrogram-based models. To the best of our knowledge, this study is the first attempt at comparing both frequency and self-supervised feature extractions in terms of adversarial robustness.

B. ADVERSARIAL ATTACK

Adversarial attacks aim to generate subtle perturbations that lead to incorrect classification by deep learning models. Given a benign input $\boldsymbol{x}$, an adversarial example $\boldsymbol{x}' = \boldsymbol{x} + \boldsymbol{\delta}$ is generated by the following maximization:

$$\max_{\|\boldsymbol{\delta}\| \leq \epsilon} \ell(\boldsymbol{x} + \boldsymbol{\delta}, \boldsymbol{y}), \tag{1}$$

where ϵ is the maximum perturbation size and $\ell(\cdot)$ is a loss function. In speech classification tasks, we generally use cross-entropy loss [10] combined with diverse losses such as decibel-based loss [11].

Adversarial attacks can be classified into two types: *white-box* and *black-box*. White-box attacks correspond to the case of an attacker accessing all of a target model's information, including its structure and parameters [22], [23]. Currently, the most powerful white-box attacks generally use the gradient information to maximize the loss. For example, projected gradient descent (PGD), which is considered as a simple and powerful adversarial attack, optimizes the perturbation $\boldsymbol{\delta}$ with a number of steps as follows:

$$\boldsymbol{\delta}_{t+1} = \Pi_{\|\boldsymbol{\delta}_t\| \leq \epsilon}(\boldsymbol{\delta}_t + \alpha \cdot \text{sign}(\nabla_{\boldsymbol{\delta}} \ell(\boldsymbol{x} + \boldsymbol{\delta}, \boldsymbol{y}))), \tag{2}$$

where Π is a projection and α is a step size. In classification tasks, the cross-entropy loss is used as a default loss ℓ. However, as Carlini et al. [23] proposed, a customized loss can be used to generate adversarial examples:

$$\ell(\boldsymbol{x} + \boldsymbol{\delta}, \boldsymbol{y}) = -\max(\max_{i \neq y} f(\boldsymbol{x} + \boldsymbol{\delta})_i - f(\boldsymbol{x} + \boldsymbol{\delta})_y, 0), \tag{3}$$

where $f(\cdot)_i$ is the i-th component of $f(\cdot)$. The attack that uses the above loss function called the Carlini and Wagner attack (CW). Both attacks are mainly applied in the vision domain, but numerous studies have demonstrated that they also show a high attack success rate in the audio domain [10], [11].

Black-box attacks are those in which an attacker only has access to the publicly available information of a target model, such as its predictions. Due to the limited access, black-box attacks generally show a lower attack success rate compared to white-box attacks, under the same computational cost. To overcome a low attack success rate under the black-box setting, a transfer attack uses another model that has been trained on similar tasks [24]. To be specific, transfer attacks first generate adversarial examples from the model with full knowledge of its structure and parameters. Then, these adversarial examples are fed to the target model. Prior studies have found that transfer attacks generally achieve a higher attack success rate over diverse tasks [8], [24], [25], [26]. In the audio domain, both white-box and black-box attacks exhibit high performance [10], [11].

C. ADVERSARIAL DEFENSE

To defend models against adversarial attacks, numerous techniques have been proposed, such as adversarial training [22], [27], using randomized neural networks [28], [29], and detection-based defense. Among them, *detection-based defense* aims to distinguish between benign and adversarial examples [30], [31], [32]. Detection-based defenses depend on malicious examples typically having different distributions compared to those of benign examples.

Several studies also have focused on detection-based defense to protect speech recognition systems. To mitigate the effect of adversarial perturbation, most defenses add noise to the input waveform [33], [34] or pad noise with sound reverberation [16]. Recently, Park et al. [35] proposed a detection method called logit noising (LN), which achieves the state-of-the-art performance against adversarial attacks by adding noise to the output of an encoder. They inject noise that impacts the prediction of adversarial audio examples, but hardly changes the prediction of benign audio examples. Given the encoder f that outputs the intermediate results (or logits) and the decoder g, the detection algorithm classifies an input audio $\boldsymbol{x}$ as an adversarial audio example when the following condition is satisfied:

$$\mathbb{E}_{\eta \sim \mathcal{N}(0, \sigma^2 I)}[\|g(f(\boldsymbol{x})) - g(f(\boldsymbol{x}) + \eta)\|] > \tau, \tag{4}$$

where $\mathcal{N}(0, \sigma^2 I)$ is a zero-centered normal distribution with a standard deviation σ and τ is a threshold. By doing this, adversarial audio examples that are highly sensitive to the additional noise in the latent space can be distinguished from benign audio examples, which are more robust to the same noise level.

Jayashankar et al. [15] also proposed a detection method that utilizes the high sensitivity of adversarial audio examples, but works in a different way. They manipulated the dropout rate p of the model to detect adversarial audio examples by measuring uncertainty. The detection algorithm classifies an input audio $\boldsymbol{x}$ as an adversarial audio example when the following condition is satisfied:

$$\mathbb{E}_p[\|g(f(\boldsymbol{x})) - g_p(f_p(\boldsymbol{x}))\|] > \tau, \tag{5}$$

where both encoder f_p and decoder g_p depend on p. Practically, the medoid value of $g_p(f_p(\boldsymbol{x}))$ is used as $g(f(\boldsymbol{x}))$. In addition, Jayashankar et al. [15] trained simple classifiers, such as support vector machine and decision tree, on the distance distribution $\|g(f(\boldsymbol{x})) - g_p(f_p(\boldsymbol{x}))\|$ and achieved high detection accuracy for a sufficiently small p.

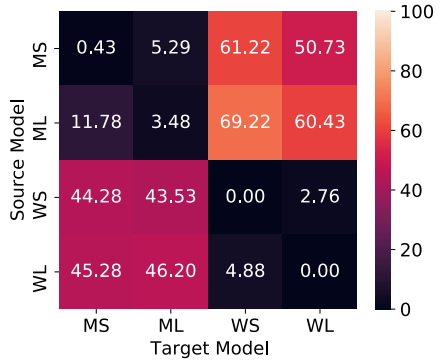


FIGURE 3. Robust accuracy (%) against adversarial audio examples generated from Mel-spectrogram-based models (MS and ML) and Wav2vec-based models (WS and WL). Source models are in rows and target models are in columns.

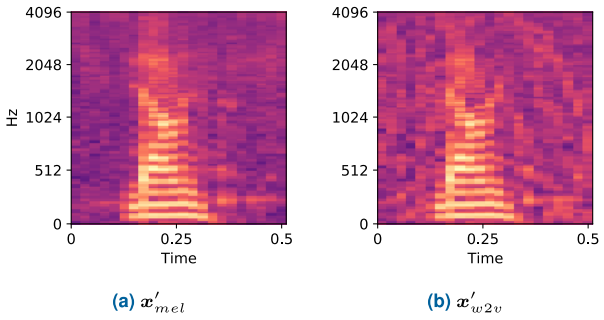


FIGURE 4. Mel-spectrogram features of adversarial audio examples generated from Model-ML (denoted as x'_{mel}) and Model-WL (denoted as x'_{w2v}). Best viewed in color.

We argue that the aforementioned works can be integrated to a detection framework that uses the following condition:

$$\mathbb{E}_{g',f'}[\|g(f(\boldsymbol{x})) - g'(f'(\boldsymbol{x}))\|] > \tau. \tag{6}$$

For example, $g' = g$ and $f'(\boldsymbol{x}) = f(\boldsymbol{x}) + \eta$ would be logit noising [35]. In this work, we focus on the observation that diverse feature extractions can improve detection performance. In other words, $f' \in \mathcal{F}$ where $\mathcal{F}$ is a set of possible distinct feature extractions. In our experiments, we demonstrate that the proposed method shows higher detection performance than existing methods by diversifying features.

III. METHODOLOGY

In this section, we first analyze the adversarial robustness of Mel-spectrogram and Wav2vec-based models. Then, we demonstrate that the adversarial audio examples generated from each feature extraction exhibit different characteristics, and thus they can be distinguished from each other. Based on this observation, we propose a new detection method that determines whether an input audio example is an adversarial audio example by diversifying feature extractions.

A. EXPLORING DIVERSE FEATURE EXTRACTIONS

To investigate adversarial robustness of different feature extractions to adversarial attacks, we train models using Mel-spectrogram and Wav2vec on the Speech Commands dataset [36]. For each type of feature extraction, we assign two neural networks with a small and large number of parameters as a classifier. In total, we obtain four different models: Mel-spectrogram + Small classifier (Model-MS), Mel-spectrogram + Large classifier (Model-ML), Wav2vec + Small classifier (Model-WS), and Wav2vec + Large classifier (Model-WL). Then, we generate adversarial audio examples with the CW attack for each model (detailed settings are presented in Section IV). Note that we observe similar results for the PGD attack.

The adversarial robustness of four models against generated adversarial examples is shown in Figure 3. The robust accuracy is defined as the percentage of correctly classified audio examples out of all the audio examples that were generated by the source models. Diagonal elements indicate the robustness against white-box attacks and other elements indicate the robustness against transfer attacks. Adversarial audio examples generated by white-box attacks successfully degenerate the performance of models (near 0% accuracy). However, interestingly, adversarial audio examples generated from different models with the same feature extraction also degenerate the performance as much as white-box adversarial audio examples. In contrast, models show a high robustness against adversarial audio examples generated from different feature extractions. This result implies that adversarial audio examples are highly dependent on feature extraction.

Motivated by Figure 3, we further visualize the Mel-spectrogram of adversarial audio examples generated from different feature extractions. In Figure 4, we present the Mel-spectrograms of randomly sampled $\boldsymbol{x}'_{mel}$ and $\boldsymbol{x}'_{w2v}$. $\boldsymbol{x}'_{w2v}$ exhibits a significantly noisy Mel-spectrogram, which is easily distinguished from $\boldsymbol{x}'_{mel}$ even to human eyes.

To further verify the difference between adversarial audio examples generated from different feature extractions, we randomly sampled 100 audio examples of $\boldsymbol{x}$, $\boldsymbol{x}'_{mel}$, and $\boldsymbol{x}'_{w2v}$ and scatter the results of principal component analysis (PCA). PC1 and PC2 denote the two principal components with the largest variance. As shown in Figure 5a, $\boldsymbol{x}'_{w2v}$ are easily distinguishable from $\boldsymbol{x}$ and $\boldsymbol{x}'_{mel}$. Similarly, $\boldsymbol{x}'_{mel}$ are easily distinguishable from $\boldsymbol{x}$ and $\boldsymbol{x}'_{w2v}$ in Figure 5b. In other words, adversarial audio examples generated from a specific feature extraction can be detected by other feature extractions.

In addition, we numerically verify the difference between features by conducting a simple experiment with a support vector machine (SVM) with RBF kernels on generated adversarial audio examples. To be specific, we first feed $\boldsymbol{x}'_{mel}$ and $\boldsymbol{x}'_{w2v}$ into other feature extractions that were not used during the attack process, i.e., Wav2vec and Mel-spectrogram, respectively. Then, we train SVM to classify benign features and generated features. For the train and test sets, we randomly split datasets into 8:2. The classification performance is summarized in Table 1. As shown in the table,

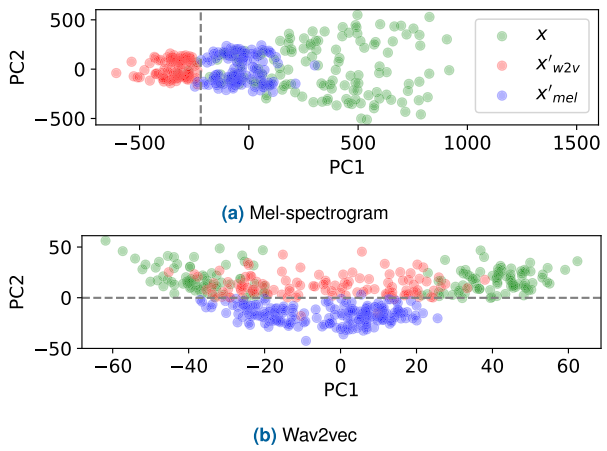


FIGURE 5. Principal component analysis (PCA) results with different feature extractions. (a) PCA result with Mel-spectrogram feature extraction. x'_{w2v} are easily distinguished from x and x'_{mel}. (b) PCA result with Wav2vec feature extraction. x'_{mel} are easily distinguished from x and x'_{w2v}.

TABLE 1. Classification performance on benign and adversarial audio examples generated from different feature-based models using an SVM.

	Accuracy	Precision	Recall	F1
Mel-spectrogram+SVM	0.92	0.90	0.93	0.92
Wav2vec+SVM	0.80	0.81	0.79	0.80

a simple SVM is enough to distinguish adversarial features from benign features. This is consistent with the result in Figure 5, which implies that benign and adversarial features are linearly separable for carefully selected components with distinct feature extractions.

We believe that this phenomenon can be related to the fact that adversarial audio examples inherit the characteristics of the feature extraction where they were generated. It can also be related to the adversarial attacks' overfitting problem [37], which demonstrate adversarial examples sometimes overfit to the source model so that they show a low transferability to the target model. In summary, the above results demonstrate that diverse feature extractions can provide benefits for detecting adversarial audio examples.

B. ENSEMBLE FEATURE EXTRACTIONS FOR ADVERSARIAL DETECTION

In the previous subsection, we verified that diverse feature extraction brings the benefits to distinguish adversarial audio examples from benign examples. However, in real-world applications, a detection method is usually forced to distinguish unseen adversarial audio examples from benign examples. In order to overcome this limitation, we propose a method to use probability information with distinct feature-based models.

First, we argue that if adversarial audio examples $\boldsymbol{x}'_{mel}$ generated from a Mel-spectrogram-based model are fed to a Wav2vec-base model, then the output probability would be different from the output probability of the Mel-spectrogram-based model, and vice-versa. This is consistent with Figure 3, which shows a high robust accuracy against adversarial audio examples generated from the different types of feature extraction. To rigorously verify this phenomenon, we measure the predictions of models against the generated adversarial audio examples ($\boldsymbol{x}'_{mel}$ and $\boldsymbol{x}'_{w2v}$) for each model.

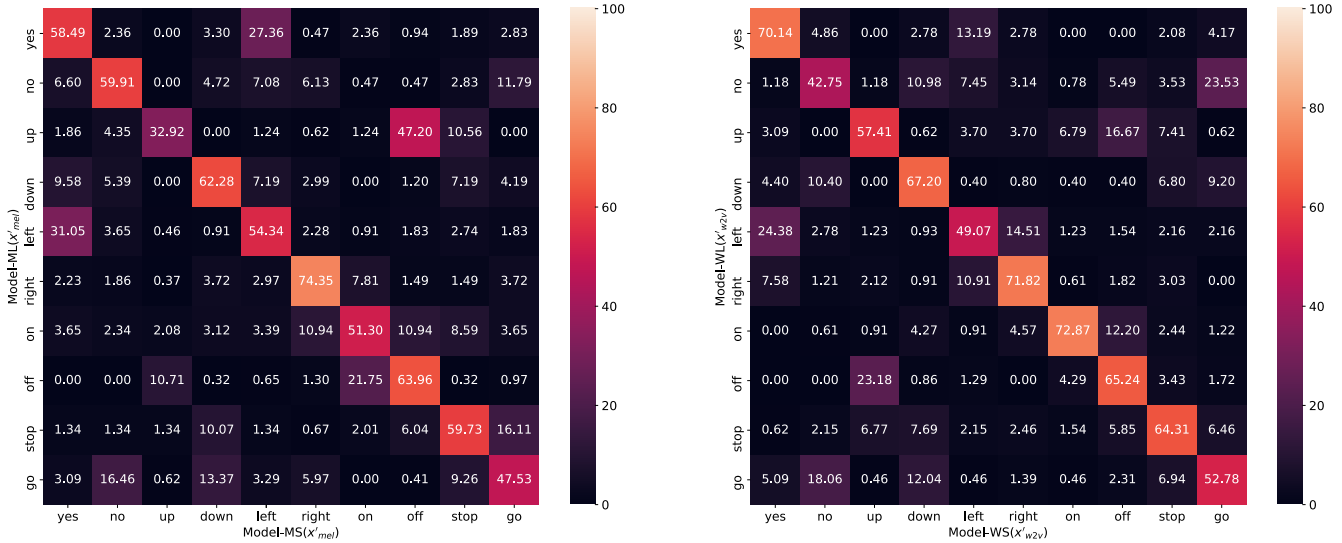
Figures 6 and 7 illustrate heatmaps of the *prediction matrix*, where each row of the matrix represents the predictions of a source model while each column represents the predictions of a target model, similar to a confusion matrix. Note that the sum of elements in each row equals to 100%.

In Figure 6, we plot the percentage of $\boldsymbol{x}'_{mel}$ and $\boldsymbol{x}'_{w2v}$ for the same feature-based models, i.e., (Model-ML and Model-MS) and (Model-WL and Model-WS). Most of the diagonal elements of the prediction matrix show the value over 50%, which implies that they highly tend to share the same probability vectors for adversarial audio examples.

On the other hand, in Figure 7, we illustrate the same prediction matrix for different feature based-models, i.e., (Model-ML and Model-WL) and (Model-WL and Model-ML). In Figure 7a, the majority of the original prediction of Model-ML, where $\boldsymbol{x}'_{mel}$ is generated, does not maintain when $\boldsymbol{x}'_{mel}$ is fed to Model-WL. In Figure 7, $\boldsymbol{x}'_{w2v}$ is more likely to output a similar prediction than $\boldsymbol{x}'_{mel}$, but all diagonal elements show values less than that of Figure 6b. We note that similar results are observed with other combinations such as Model-MS and Model-WS, which indicates this phenomenon is generally observed regardless of model structures and the number of parameters. Thus, the adversarial audio examples highly tend to have different probability vectors for different feature-based models.

To push further, we measure the output probability difference with benign and adversarial audio examples. Specifically, we first obtain the output probability vector of the audio examples with Model-ML and Model-WL, then calculate the distance between the two output probability vectors. Here, we use the L_1 norm distance as a default, but we note that similar results are observed with L_2 norm distance. As shown in Figure 9a, for benign audio examples, the probability distance is close to zero, which implies that the probability vectors obtained from Model-ML and Model-WL are similar. However, for adversarial audio examples, most of the examples show probability distances over 0.1, which is extremely higher than that of benign audio examples. Thus, due to the variability of adversarial examples with respect to feature extractions, we can detect the audio adversarial examples with their probability information.

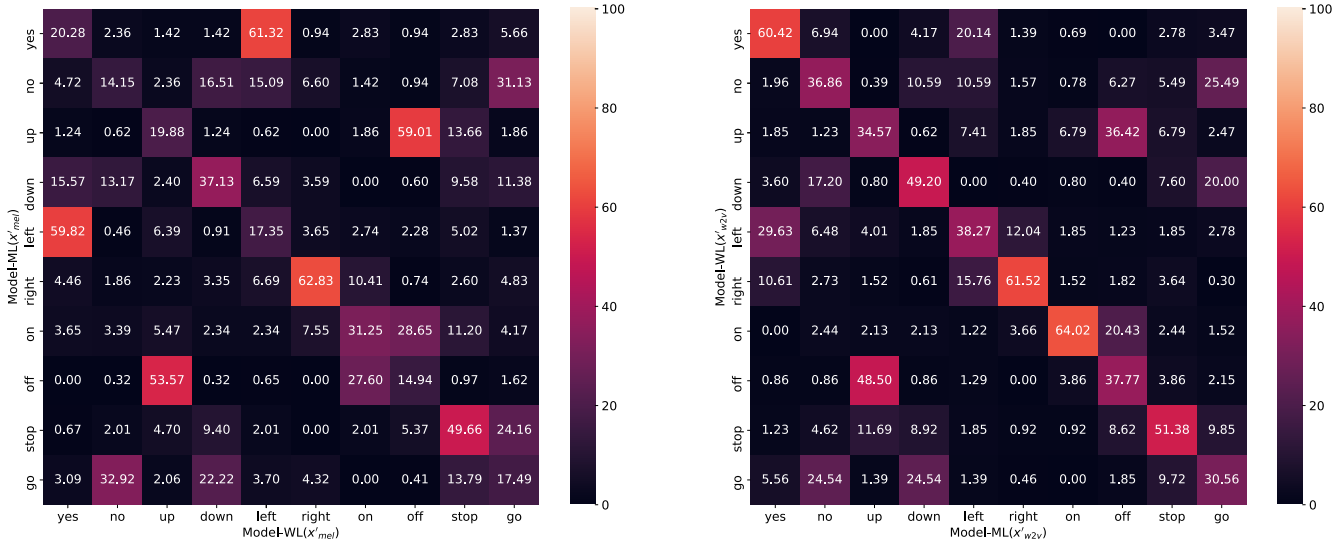
Based on the observation, we propose a new detection method, Ensemble Feature-based Detection (EFD), as shown in Figure 8. Let us denote a original feature extraction f and a classifier g as same as Equation (6). Then, an input audio example $\bar{\boldsymbol{x}}$ is initially fed to f and g to obtain the probability vector $\boldsymbol{p}$. In the proposed method, we simultaneously obtain the probability vector $\boldsymbol{p}'$ from f' and g', where f' is a distinct feature extraction from f so that it captures the variability of adversarial examples. Then, the distance between $\boldsymbol{p}$ and $\boldsymbol{p}'$ is calculated to determine whether the given audio example $\bar{\boldsymbol{x}}$



(a) x'_{mel} (Source: Model-ML, Target: Model-MS)

(b) x'_{w2v} (Source: Model-WL, Target: Model-WS)

FIGURE 6. Prediction matrix of x'_{mel} and x'_{w2v} for same feature-based models (in percentage). Best viewed in color.



(a) x'_{mel} (Source: Model-ML, Target: Model-WL)

(b) x'_{w2v} (Source: Model-WL, Target: Model-ML)

FIGURE 7. Prediction matrix of x'_{mel} and x'_{w2v} for different feature-based models (in percentage). Best viewed in color.

is an adversarial example or not. If the distance between the output probability is less than a threshold τ, then the proposed method determines $\bar{x}$ as a benign example. In contrast, if the distance between the output probability exceeds the threshold τ, then the proposed method determines $\bar{x}$ as an adversarial example.

IV. EXPERIMENTS

A. EXPERIMENTAL SETUP

In this section, we conduct experiments on two popular speech classification tasks: speech commands classification on the Google Speech Commands dataset [36] and speaker recognition on the VCTK dataset [38]. The Google Speech Commands dataset includes 65000 samples in 30 classes such as "go" and "stop". The VCTK dataset includes utterances by 110 English speakers. Following [10] and [39], we extract samples from the 10 most used labels in both datasets.

1) SPEECH COMMANDS CLASSIFICATION

For Mel-spectrogram-based models, we use 32 Mel-filterbanks and the hop length is set to 512 so that Mel-spectrogram outputs a feature with the size 32 × 32.

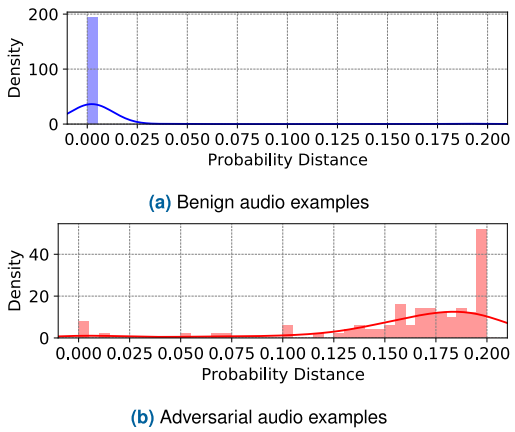


FIGURE 8. Illustration of the proposed detection method, Ensemble Feature-based Detection (EFD). The input audio example is fed to distinct feature extractions f and f'; then, based on the distance between the probability vectors and the threshold, whether the given input audio is adversarial or not.

Then, we assign two neural networks with small and large numbers of parameters as the classifier. We use DenseNet [40] as a small model (769,416 parameters) and Wide-ResNet 28-10 (WRN) [41] as a large model (36,480,188 parameters). For Wav2vec-based models, we use the pre-trained Wav2vec that outputs a feature with the size 512×98. Then, similar to the Mel-spectrogram-based models, we use a fully connected model (6,381,580 parameters) and a long short-term memory model with the self-attention model (7,886,860 parameters) as a small and large model, respectively. In total, we use two models for each feature extraction as introduced in Section III: Mel-spectrogram-based models (Model-MS and Model-ML) and Wav2vec-based models (Model-WS and Model-WL). We train all models over 70 epochs with Adam. An initial learning rate is set to 0.005 and cosine learning rate decay is used.

2) SPEAKER RECOGNITION

For Mel-spectrogram-based models, we use 30 Mel-filterbanks and the hop length is set to 512 so that Mel-spectrogram outputs a feature with the size 30×44. We use X-vector [42] for the classifier and control the number of parameters by manipulating the channels of convolution filters. The large model has 6,060,518 parameters and the small model has 3,081,190 parameters. In total, four models are used as same in the speech commands classification task. For Wav2vec models, we use the same models in the speech commands classification task. All models are trained with the same setting used in the speech commands classification task.

3) ATTACK SETTINGS

To measure the general detection performance against adversarial attacks, we considered both white- and black-box adversarial attacks. The white-box adversarial examples were generated on the same model using the same feature extraction and classifier. The black-box adversarial examples were generated on the model using different feature extractions or different classifiers. For both cases, we use the CW and PGD attacks with decibel regularization, as proposed in [11], to satisfy $dB(|\boldsymbol{x}' - \boldsymbol{x}|) \leq dB(\boldsymbol{x}) - 20$ where $dB(\boldsymbol{x}) = \max_t 20 \cdot \log_{10}(x_t)$. PyTorch [43] and Torchattacks [44] are mainly used for the experiments on six NVIDIA TITAN V GPUs and an Intel Xeon(R) Gold 6126 CPU using the Ubuntu 16.04 OS.

B. BASELINES AND METRICS

We consider the previous state-of-the-art detection methods, logit noising (LN) [35] and dropout uncertainty (DU) [15], as baselines. Logit noising is a method that distinguishes adversarial audio examples from benign audio examples by injecting noise into the logit space. Following [35], the noise is drawn from the normal distribution and added to the feature space. Dropout uncertainty is a method that determines whether an input audio example is an adversarial audio example by manipulating the dropout rate of models. Following [15], we use SVM to classify the uncertainty distribution. For the propose method (EFD), we emphasize that any model can be used as a combined model if it uses a different feature extraction from the base model due to the model-independent characteristic of the proposed framework. We basically use a large model that uses a different feature extraction as f' and g', but we note that using a small model is also enough to gain sufficient improvement in detection rates. We estimate the four different detection performance measures: accuracy, precision, recall, and F1 scores. For all methods, we perform grid searches over the threshold τ on the training set, then report the best performance that shows the highest F1 score. Each experiment is conducted with three different random seeds. We report the average and standard deviation of each detection performance measure.

C. RESULTS AGAINST WHITE-BOX ATTACKS

In Table 2, we summarize the detection performance against the white-box attacks. For most of the measures, the proposed method shows the best performance. Specifically, the proposed method outperforms other methods on the speaker recognition task. On the speech classification task, the proposed method achieves near 0.9 F1 scores for both Mel-spectrogram-based models and Wav2vec-based models. On the speaker recognition task, the proposed method outperforms other baselines, especially on Wav2vec-based models. Considering that the speaker recognition task is more complicated than the speech classification task, we believe that using both Mel-spectrogram and Wav2vec feature extractions becomes more important in terms of adversarial robustness.

D. RESULTS AGAINST BLACK-BOX ATTACKS

In Table 3, we evaluate the detection performance against the black-box attacks. In this experiment, the adversarial audio examples are generated from other models that use different feature extractions and classifiers. Similar to the white-box attack cases, the proposed method shows stable performance across all tasks and models with high detection rates. Especially, for the speaker recognition task and

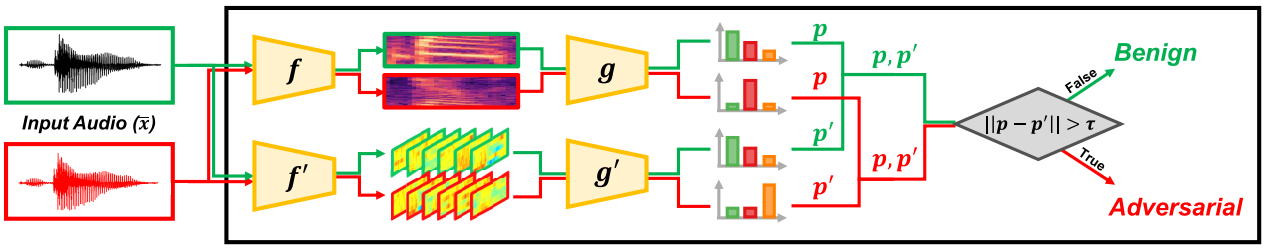


FIGURE 9. Distribution of the distance between probability vectors of Model-ML and Model-WL for 200 randomly sampled benign and adversarial audio examples.

TABLE 2. Performance comparison of detection methods against white-box attacks. Higher is better.

Task	Model	Defense	White-box			
			Accuracy	Precision	Recall	F1
Speech Classification	Mel-spectrogram -based	LN	0.87±0.004	0.86±0.013	0.88±0.010	0.87±0.002
		DU	0.75±0.002	0.78±0.006	0.70±0.007	0.73±0.002
		EFD	**0.91±0.006**	**0.92±0.008**	**0.90±0.023**	**0.91±0.007**
	Wav2vec-based	LN	0.78±0.014	0.76±0.006	0.82±0.054	0.78±0.022
		DU	0.81±0.009	0.78±0.020	0.87±0.015	0.82±0.005
		EFD	**0.87±0.010**	**0.88±0.039**	**0.88±0.029**	**0.87±0.005**
Speaker Recognition	Mel-spectrogram -based	LN	0.85±0.005	0.90±0.007	0.79±0.020	0.84±0.008
		DU	0.82±0.013	0.90±0.051	0.72±0.031	0.80±0.009
		EFD	**0.86±0.004**	**0.92±0.027**	**0.80±0.029**	**0.86±0.006**
	Wav2vec-based	LN	0.68±0.013	0.70±0.028	0.63±0.021	0.66±0.004
		DD	0.61±0.027	0.61±0.056	0.61±0.082	0.61±0.017
		EFD	**0.83±0.001**	**0.79±0.011**	**0.91±0.018**	**0.84±0.002**

TABLE 3. Performance comparison of detection methods against black-box attacks. Higher is better.

Task	Model	Defense	Black-box			
			Accuracy	Precision	Recall	F1
Speech Classification	Mel-spectrogram -based	LN	0.80±0.007	0.83±0.028	**0.75±0.022**	0.79±0.004
		DU	0.73±0.043	0.75±0.077	**0.75±0.099**	0.74±0.007
		EFD	**0.83±0.019**	**0.90±0.002**	**0.75±0.045**	**0.81±0.026**
	Wav2vec-based	LN	0.68±0.008	0.66±0.016	0.75±0.033	0.71±0.006
		DU	0.70±0.009	0.67±0.014	**0.79±0.018**	0.72±0.001
		EFD	**0.82±0.024**	**0.90±0.001**	0.68±0.043	**0.77±0.028**
Speaker Recognition	Mel-spectrogram -based	LN	0.84±0.001	0.84±0.014	**0.85±0.019**	0.84±0.004
		DU	0.85±0.003	0.87±0.009	0.84±0.016	0.85±0.001
		EFD	**0.88±0.006**	**0.95±0.019**	0.80±0.013	**0.87±0.006**
	Wav2vec-based	LN	0.58±0.019	0.56±0.020	0.74±0.040	0.64±0.010
		DD	0.59±0.016	0.58±0.033	0.69±0.040	0.63±0.004
		EFD	**0.82±0.002**	**0.78±0.000**	**0.89±0.004**	**0.83±0.002**

Wav2vec-based models, the proposed method shows over 20% improvements in accuracy compared to other baselines. Since both comparison methods [15], [35] depend on the specific feature extraction that the target model used, they hardly detect adversarial examples generated from unseen feature extractions.

V. CONCLUSION

In this work, we investigate and analyze the adversarial robustness of different feature extractions: Mel-spectrogram and Wav2vec, representing traditional and modern feature extractions in the audio domain, respectively. We discover that they yield distinct latent spaces that help us determine whether an input is an adversarial audio example. The proposed method reveals the effectiveness of using diverse feature extractions and affords high detection rates for both white-box and black-box adversarial audio examples. Our observation points toward potential directions for future research on more complicated speech challenges, such as speech-to-text tasks. Thus, in future work, we plan to

investigate the effect of utilizing diverse feature extractions in speech-to-text systems with recent models and to improve the proposed method to provide a secure speech recognition system on large-scale datasets.

REFERENCES

[1] C. Dewi, R.-C. Chen, and Y.-T. Liu, "Similar music instrument detection via deep convolution YOLO-generative adversarial network," in *Proc. IEEE 10th Int. Conf. Awareness Sci. Technol. (iCAST)*, Oct. 2019, pp. 1–6.

[2] H. Zhao, C. Wang, R. Guo, X. Rong, J. Guo, Q. Yang, L. Yang, Y. Zhao, and Y. Li, "Autonomous live working robot navigation with real-time detection and motion planning system on distribution line," *High Voltage*, vol. 7, no. 6, pp. 1204–1216, Dec. 2022.

[3] L. Deng, J. Li, J.-T. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams, Y. Gong, and A. Acero, "Recent advances in deep learning for speech research at Microsoft," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 8604–8608.

[4] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, "Deep speech: Scaling up end-to-end speech recognition," 2014, *arXiv:1412.5567*.

[5] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "Wav2vec: Unsupervised pre-training for speech recognition," 2019, *arXiv:1904.05862*.

[6] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," 2013, *arXiv:1312.6199*.

[7] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, *arXiv:1412.6572*.

[8] 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.

[9] A. Kurakin, "Adversarial examples in the physical world," in *Artificial Intelligence Safety and Security*. London, U.K.: Chapman & Hall, 2018, pp. 99–112.

[10] M. Alzantot, B. Balaji, and M. Srivastava, "Did you hear that? Adversarial examples against automatic speech recognition," 2018, *arXiv:1801.00554*.

[11] 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.

[12] F. Kreuk, Y. Adi, M. Cisse, and J. Keshet, "Fooling end-to-end speaker verification with adversarial examples," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 1962–1966.

[13] M. Cisse, Y. Adi, N. Neverova, and J. Keshet, "Houdini: Fooling deep structured prediction models," 2017, *arXiv:1707.05373*.

[14] D. Wang, L. Dong, R. Wang, D. Yan, and J. Wang, "Targeted speech adversarial example generation with generative adversarial network," *IEEE Access*, vol. 8, pp. 124503–124513, 2020.

[15] T. Jayashankar, J. L. Roux, and P. Moulin, "Detecting audio attacks on ASR systems with dropout uncertainty," 2020, *arXiv:2006.01906*.

[16] X. Du, C.-M. Pun, and Z. Zhang, "A unified framework for detecting audio adversarial examples," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 3986–3994.

[17] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen, and J. Chen, "Deep speech 2: End-to-end speech recognition in English and Mandarin," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 173–182.

[18] Y. Takashima, R. Takashima, T. Takiguchi, and Y. Ariki, "Knowledge transferability between the speech data of persons with dysarthria speaking different languages for dysarthric speech recognition," *IEEE Access*, vol. 7, pp. 164320–164326, 2019.

[19] J.-W. Kim, H. Yoon, and H.-Y. Jung, "Linguistic-coupled age-to-age voice translation to improve speech recognition performance in real environments," *IEEE Access*, vol. 9, pp. 136476–136486, 2021.

[20] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 12449–12460.

[21] H. Wu, B. Zheng, X. Li, X. Wu, H.-Y. Lee, and H. Meng, "Characterizing the adversarial vulnerability of speech self-supervised learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 3164–3168.

[22] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2017, *arXiv:1706.06083*.

[23] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2017, pp. 39–57.

[24] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in machine learning: From phenomena to black-box attacks using adversarial samples," 2016, *arXiv:1605.07277*.

[25] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," 2016, *arXiv:1611.02770*.

[26] S. Cheng, Y. Dong, T. Pang, H. Su, and J. Zhu, "Improving black-box adversarial attacks with a transfer-based prior," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 10934–10944.

[27] H. Kim, W. Lee, and J. Lee, "Understanding catastrophic overfitting in single-step adversarial training," 2020, *arXiv:2010.01799*.

[28] V. Sehwag, S. Wang, P. Mittal, and S. Jana, "Hydra: Pruning adversarially robust neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 19655–19666.

[29] S. Lee, H. Kim, and J. Lee, "GradDiv: Adversarial robustness of randomized neural networks via gradient diversity regularization," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Apr. 21, 2022, doi: 10.1109/TPAMI.2022.3169217.

[30] K. Grosse, P. Manoharan, N. Papernot, M. Backes, and P. McDaniel, "On the (statistical) detection of adversarial examples," 2017, *arXiv:1702.06280*.

[31] J. Liu, W. Zhang, Y. Zhang, D. Hou, Y. Liu, H. Zha, and N. Yu, "Detection based defense against adversarial examples from the steganalysis point of view," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4825–4834.

[32] X. Zhang, Y. Zhou, S. Pei, J. Zhuge, and J. Chen, "Adversarial examples detection for XSS attacks based on generative adversarial networks," *IEEE Access*, vol. 8, pp. 10989–10996, 2020.

[33] K. Rajaratnam and J. Kalita, "Noise flooding for detecting audio adversarial examples against automatic speech recognition," in *Proc. IEEE Int. Symp. Signal Process. Inf. Technol. (ISSPIT)*, Dec. 2018, pp. 197–201.

[34] H. Kwon, H. Yoon, and K.-W. Park, "POSTER: Detecting audio adversarial example through audio modification," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Nov. 2019, pp. 2521–2523.

[35] N. Park, S. Ji, and J. Kim, "Detecting audio adversarial examples with logit noising," in *Proc. Annu. Comput. Secur. Appl. Conf.*, Dec. 2021, pp. 586–595.

[36] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," 2018, *arXiv:1804.03209*.

[37] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, and A. L. Yuille, "Improving transferability of adversarial examples with input diversity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2730–2739.

[38] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," Centre Speech Technol. Res. (CSTR), Univ. Edinburgh, Edinburgh, Scotland, 2019. [Online]. Available: https://datashare.ed.ac.uk/handle/10283/3443, doi: 10.7488/DS/2645.

[39] Z. Li, Y. Wu, J. Liu, Y. Chen, and B. Yuan, "AdvPulse: Universal, synchronization-free, and targeted audio adversarial attacks via subsecond perturbations," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2020, pp. 1121–1134.

[40] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

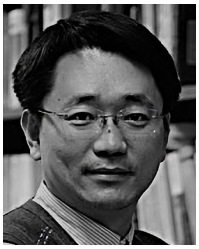
[42] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5329–5333.

[43] A. Paszke, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8026–8037.

[44] H. Kim, "Torchattacks: A PyTorch repository for adversarial attacks," 2020, *arXiv:2010.01950*.

YUJIN CHOI received the B.S. degree in industrial engineering and mathematics from Yonsei University, in 2021. She is currently pursuing the Ph.D. degree in industrial engineering with Seoul National University, Seoul, South Korea. Her research interests include adversarial robustness, privacy, and security.

JINSEONG PARK received the B.S. degree in industrial and management engineering from the Pohang University of Science and Technology, in 2020, and the M.S. degree in industrial engineering from Seoul National University, Seoul, South Korea, in 2022, where he is currently pursuing the Ph.D. degree with the Department of Industrial Engineering. His research interests include safety, privacy, and generalization of machine learning and deep learning.



JAEWOOK LEE received the B.S. degree in mathematics from Seoul National University, Seoul, South Korea, in 1993, and the Ph.D. degree in applied mathematics from Cornell University, in 1999. He is currently a Professor with the Department of Industrial Engineering, Seoul National University. His research interests include machine learning, neural networks, and global optimization and their applications to data mining and financial engineering.

HOKI KIM received the B.S. degree in industrial engineering from Seoul National University, Seoul, South Korea, in 2018, where he is currently pursuing the Ph.D. degree with the Department of Industrial Engineering. His research interests include robustness and generalization of machine learning and deep learning models.