Detecting and Defending Against Adversarial Attacks on Automatic Speech Recognition via Diffusion Models

Nikolai L Kühne, Astrid H F Kitchen, Marie S Jensen, Mikkel S L Brøndt, Martin Gonzalez, Christophe Biscio, Zheng-Hua Tan

Detecting and Defending Against Adversarial Attacks on Automatic Speech Recognition via Diffusion Models

Nikolai L. Kühne*°, Astrid H. F. Kitchen*°, Marie S. Jensen*°, Mikkel S. L. Brøndt*°, Martin Gonzalez†, Christophe Biscio*, and Zheng-Hua Tan*
*Aalborg University, Denmark
Email: nlk@es.aau.dk, {akitch20, marije19, mbrand20}@student.aau.dk, christophe@math.aau.dk, zt@es.aau.dk
†IRT SystemX, France
Email: martin.gonzalez@irt-systemx.fr

***Abstract*—Automatic speech recognition (ASR) systems are known to be vulnerable to adversarial attacks. This paper addresses detection and defence against targeted white-box attacks on speech signals for ASR systems. While existing work has utilised diffusion models (DMs) to purify adversarial examples, achieving state-of-the-art results in keyword spotting tasks, their effectiveness for more complex tasks such as sentence-level ASR remains unexplored. Additionally, the impact of the number of forward diffusion steps on performance is not well understood. In this paper, we systematically investigate the use of DMs for defending against adversarial attacks on sentences and examine the effect of varying forward diffusion steps. Through comprehensive experiments on the Mozilla Common Voice dataset, we demonstrate that two forward diffusion steps can completely defend against adversarial attacks on sentences. Moreover, we introduce a novel, training-free approach for detecting adversarial attacks by leveraging a pre-trained DM. Our experimental results show that this method can detect adversarial attacks with high accuracy.**

***Index Terms*—Automatic speech recognition, adversarial attacks, diffusion models.**

I. Introduction

Automatic speech recognition (ASR) systems, like other machine learning models, are vulnerable to adversarial attacks. These attacks involve adding subtle, optimised perturbations to the original input to deceive the ASR system, while remaining imperceptible to humans [1]–[10]. The consequences of such attacks are severe, potentially leading to compromised security systems, unauthorised purchases, and theft of sensitive information stored on devices controlled by voice assistants. Given the widespread use of ASR systems in mobile phones, smart devices, and vehicles, the need to detect and defend against adversarial attacks has become increasingly urgent [11].

Existing defensive methods, including input transformation-based defences and distillation [12]–[15], reduce adversarial attack success rate but are less effective against strong (e.g., white-box (WB)) attacks [16], [17] and ineffective against adaptive attacks [12]. Adversarial training [18], considered the most effective defence according to [19], still leaves models vulnerable to attacks not encountered during training.

° These authors contributed equally to this work.

In this work, we address key issues by leveraging a pre-trained diffusion model (DM) [20]–[22] to detect and defend against targeted WB attacks on ASR systems. Targeted attacks aim to trick the ASR system into misclassifying inputs as a specific, incorrect class, with WB attacks having full access to the parameters of the target model.

By exploring the inherent denoising property of denoising diffusion probabilistic models (DDPMs) [21], an adversarial purification-based method for audio was proposed in [19] based on the pre-trained DDPM from [23]. This method proved to be the most effective against WB attacks on keywords compared to other cutting-edge defence methods. However, these methods have not been tested on sentences, which ASR systems typically process. This gap motivates our approach: applying diffusion models to sentences and exploring the impact of varying forward diffusion steps. Each step adds noise to override the adversarial perturbations, with the reverse diffusion process generating purified data from the noisy inputs, thereby enhancing the robustness of ASR systems against adversarial attacks.

Another approach for safeguarding ASR systems is detecting adversarial attacks. There are two main methods: one involves constructing a specialised classifier, while the other can generalise to a wider range of threats without needing a specialised classifier. A method in the first category [24] uses a convolutional neural network with small kernels to identify subtle perturbations in adversarial examples. This was enhanced in [25] by incorporating filter bank-based features for better detection. However, these methods are less effective against attacks dissimilar to those used in training. This motivates the second category of methods. These works either leverages dependencies among neighbouring sounds in audio sequences [12] or common signal processing transformations [26] for binary classification, achieving state-of-the-art results on strong adaptive attacks. Another approach [27] uses ASR classification scores to detect attacks without model training, but it fails when the classification scores are unknown. All these methods [12], [26], [27] were tested on only 100 samples from the Mozilla Common Voice (MCV) dataset [28]. We in-

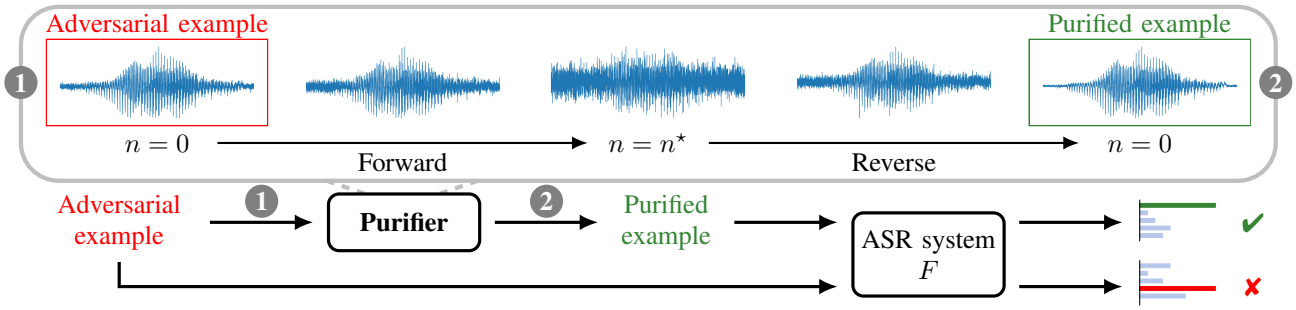

Fig. 1. The architecture of the whole speech system protected by the **Purifier**. The **Purifier** first adds noise to the adversarial example via the forward diffusion process and then runs the reverse process to obtain the purified example. Subsequently, the purified example is fed into the ASR system to get predictions. Without the **Purifier**, the adversarial example is fed into the ASR system directly. Inspired by [19].

troduce a novel approach using DMs for detecting adversarial attacks. Our method compares ASR system outputs of non-purified and purified speech signals, tested on a substantially larger numbers of samples.

Our contributions are multifold. First, we explore using a pre-trained DDPM to defend against adversarial attacks on audio at the sentence level and analyse its behaviour. Next, we systematically study the impact of varying forward diffusion steps on ASR performance for both clean speech and adversarial examples. Extensive experiments on a subset of the MCV dataset [28] show that while the pre-trained DDPM can completely defend against adversarial attacks, ASR performance on clean speech degrades. We also found that two diffusion steps yield a 100% defence success rate. Finally, we propose a novel purification-based detection method for adversarial attacks using the same pre-trained DDPM. Our method, tested on a larger subset of the MCV dataset than those used in the literature, can be run on consumer-grade hardware. Both our code and generated datasets are publicly available.[1] The system architecture is illustrated in Fig. 1.

II. Methods

A. Adversarial Attacking Method

The Carlini & Wagner (C&W) attacking method [16] is chosen for its widespread adoption as one of the most successful targeted WB attacking methods for audio. It achieves 100% adversarial attack success rate on targeted attacks on a subset of the MCV dataset assuming the target is theoretically reachable [16]. In the method, the added adversarial perturbation $\boldsymbol{\delta}$ is optimised by iteratively solving the optimisation problem

$$\underset{\boldsymbol{\delta}}{\text{minimise}} \ \ \|\boldsymbol{\delta}\|_2^2 + c \cdot \ell(\boldsymbol{x} + \boldsymbol{\delta}, \boldsymbol{t}) \tag{1}$$

$$\text{such that } dB(\boldsymbol{\delta}) - dB(\boldsymbol{x}) \leq \tau, \tag{2}$$

where c is a regularisation term, $\ell(\cdot)$ is the connectionist temporal classification loss function [29], $\boldsymbol{x} \in \mathbb{R}^n$ is the original example, and $\boldsymbol{t}$ is the desired target phrase. Furthermore, the constant τ is initially sufficiently large to ensure a partial solution $\boldsymbol{\delta}^*$ exists, and then in each iteration τ is reduced until no solution can be found.

[1] https://github.com/Kyhne/Detecting-and-Defending-Against-Adversarial-Attacks

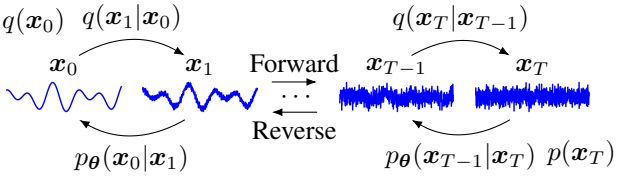

Fig. 2. The forward and reverse diffusion process in DMs.

B. Denoising Diffusion Probabilistic Models

For DMs, this work uses the DDPM framework from [21]. A DM consists of a forward diffusion process and a reverse diffusion process, defined by Markov chains, as shown in Fig. 2. The forward process gradually adds noise to the input data until the distribution of the noisy data approximately equals a standard Gaussian distribution. The reverse process is parameterised by a deep neural network that takes the approximately standard Gaussian noise as input and gradually denoises the data to recover clean data.

Formally, let $\boldsymbol{x}_0 \in \mathbb{R}^n$ and $q(\boldsymbol{x}_0)$ be an unknown data distribution. For each data point $\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0)$, a forward Markov chain is formed such that

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}\big(\boldsymbol{x}_t\big|\sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t\boldsymbol{I}\big), \ 1 \leq t \leq T, \tag{3}$$

based on a pre-determined noise variance schedule $\{\beta_t\}_{t=1}^T$, where $0 < \beta_t < 1$. Using the reparameterisation trick [30] with $\alpha_t = 1 - \beta_t$ and $\overline{\alpha}_t = \prod_{i=0}^t \alpha_i$ results in

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}\big(\boldsymbol{x}_t\big|\sqrt{\overline{\alpha}_t}\boldsymbol{x}_0, (1-\overline{\alpha}_t)\boldsymbol{I}\big). \tag{4}$$

The sequence $\{\beta_t\}_{t=1}^T$ is chosen such that $\overline{\alpha}_T \approx 0$, which results in $q(\boldsymbol{x}_T|\boldsymbol{x}_0) \approx \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, and thus $q(\boldsymbol{x}_T) \approx \mathcal{N}(\boldsymbol{x}_T|\boldsymbol{0}, \boldsymbol{I})$. Since the reverse diffusion process is intractable, the denoising process $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ is approximated by a learnable Markov chain with parameters $\boldsymbol{\theta}$ defined by

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}\big(\boldsymbol{x}_{t-1}|\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t), \boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)\big) \tag{5}$$

given a prior distribution $p(\boldsymbol{x}_T) = \mathcal{N}(\boldsymbol{x}_T|\boldsymbol{0}, \boldsymbol{I})$. Specifically,

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)\right), \tag{6}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) = \frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t}\beta_t\boldsymbol{I}, \tag{7}$$

where $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}} : \mathbb{R}^n \times \mathbb{N} \to \mathbb{R}^n$ is a deep neural network predicting the noise $\boldsymbol{\epsilon}_0 \sim \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \boldsymbol{I})$ in the forward process. The choice of $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$ is based on [19] as it delivers superior results for adversarial purification on audio.

In this work, we use an adversarial purification-based method for defence denoted as **Purifier**, depicted in Fig. 1. The **Purifier** is based on the plug-and-play method AudioPure [19], where we omit the Wave2Mel module that extracts Mel spectrograms as we focus solely on waveform ASR systems.

C. Method for Defending Against Adversarial Attacks

For defence, a waveform input is passed through the **Purifier** and a pre-trained ASR system, sequentially. The **Purifier** is based on a pre-trained DDPM: DiffWave [23]. DiffWave is a 36-layer, 6.91M parameter residual neural network [31] and uses a bidirectional dilated convolution architecture. The kernel size is 3, the dilation cycle is $[1, 2, 4, \ldots, 2048]$, and the number of residual channels is $C = 256$. Furthermore, the number of forward steps is $T = 200$, and the noise schedule is linearly spaced for $0.0001 \leq \beta_t \leq 0.02$.

Let $\boldsymbol{x}_{\text{adv}} \in \mathbb{R}^n$ be the adversarial waveform input. To override the adversarial perturbations, the **Purifier** adds noise to $\boldsymbol{x}_{\text{adv}}$ via the forward diffusion process. From the T forward diffusion steps, only the first $n^\star \in \{1, 2, \ldots, T\}$ steps are used. This prevents excessive degradation of the original waveform, which could hinder recovery and lead to misclassifications. The reverse diffusion process then aims to reconstruct the clean signal from the noisy input. The grey box in Fig. 1 depicts this process. Formally, the **Purifier** takes $\boldsymbol{x}_{\text{adv}}$ and $n^\star$ as inputs, and returns the purified-speech signal leading to the function **Purifier** $: \mathbb{R}^n \times \mathbb{N} \to \mathbb{R}^n$. Next, the ASR system $F : \mathbb{R}^n \to \mathbb{R}^d$ is applied to the purified-speech signal, where d is the length of the output. Finally, the **Purifier** and the ASR system are combined into a defended speech system **SS** $: \mathbb{R}^n \times \mathbb{N} \to \mathbb{R}^d$ shown in Fig. 1 and given by

$$\textbf{SS}\left(\boldsymbol{x}_{\text{adv}}, n^\star\right) = F\left(\textbf{Purifier}\left(\boldsymbol{x}_{\text{adv}}, n^\star\right)\right). \tag{8}$$

D. Method for Detecting Adversarial Attacks

We propose a detection method using a binary classifier that labels inputs as either adversarial or benign. To classify an input sentence $\boldsymbol{x}_{\text{in}}$, the character error rate (CER) is calculated between the non-purified speech ASR output $F(\boldsymbol{x}_{\text{in}})$ and the purified speech ASR output $\mathbf{SS}(\boldsymbol{x}_{\text{in}}, n^\star)$. If $\text{CER}(\mathbf{SS}(\boldsymbol{x}_{\text{in}}, n^\star), F(\boldsymbol{x}_{\text{in}})) > \Omega$ for a pre-determined threshold Ω and forward diffusion steps $n^\star$, $\boldsymbol{x}_{\text{in}}$ is classified as adversarial, otherwise as benign. The threshold Ω and forward diffusion steps $n^\star$ are found through a grid search systematically exploring different values of hyperparameters. CER is used as opposed to word error rate (WER), since preliminary experiments showed that CER yields better results.

III. Experiments and Results

A. Defending Against Adversarial Attacks

We experiment with adversarial defence and purification for sentences with varying forward diffusion steps. In all experiments $n^\star \in \{1, 2, 3, 4, 5\}$. The dataset and models were selected due to the high adversarial attack success rate on the ASR system and dataset [16].

Datasets. From the MCV dataset [28], 300 short (S), 300 medium (M), and 300 long (L) English sentences of length 1-2, 3-4, and 6-7 seconds, respectively, are chosen. Only audio files containing more than 68% speech have been chosen using the open-source robust Voice Activity Detection (rVAD) algorithm [32]. All sentences are attacked with the same WB attack using the same target for sentences of the same length:

- **Short target:** *open all doors*.
- **Medium target:** *switch off internet connection*.
- **Long target:** *i need a reservation for sixteen people at the seafood restaurant down the street*.

Attack Method. The C&W [16] method is used, where the maximum amount of iterations is 5000 and the learning rate is 10 as in [16].

Models. The unconditional version of DiffWave [23] with the officially provided pre-trained checkpoint is utilised as the defensive waveform purifier. A pre-trained end-to-end Deep Speech system[2] [33] is used as an ASR system.

Evaluation metrics. The adversarial attack success rate is computed as the percentage of times the ASR system transcribes exactly the target of the attack.

The ASR performance is represented as $1 - \text{CER}$, where CER is calculated between the transcribed output and the ground truth label. CER is used instead of WER, since error rate at the character level provides higher granularity.

B. Detecting Adversarial Attacks

In terms of **datasets, attack method, and models**, adversarial detection is performed on sentences using the same configurations as detailed in Section III-A.

Attacking all 900 benign sentences yields 900 adversarial sentences. As such, we generate 300 benign and 300 adversarial sentences for each sentence lengths resulting in a dataset containing 1800 sentences. For each sentence length, we split the data into two parts: the first 10% for determining the threshold Ω and the number of forward diffusion steps $n^\star$, and the next 90% for testing, each part containing 50% adversarial and 50% benign sentences. For each $n^\star \in \{1, 2, 3, 4, 5\}$, a grid search is performed to find the Ω that maximises the adversarial detection accuracy across all sentence lengths, leading to the global hyperparameters $(n^\star, \Omega)$.

TABLE I
AVERAGE RUNTIME PER SENTENCE LENGTH IN SECONDS AVERAGED OVER 300 EXAMPLES.

	$n^\star$	S	M	L
Purifier	1	0.51	0.88	1.67
Purifier	5	1.44	2.68	4.84
Deep Speech	–	0.68	1.30	2.22
Detection	2	2.17	4.19	8.40

[2]https://github.com/mozilla/DeepSpeech/releases/tag/v0.9.3, 2020

TABLE II
ASR PERFORMANCE MEASURED AT CHARACTER LEVEL.

$n^\star$	Clean-S	WB-S	Clean-M	WB-M	Clean-L	WB-L
0	**78.00**	19.41	**86.01**	20.58	**86.07**	24.67
1	72.10 ± 0.66	42.62 ± 0.34	75.71 ± 0.45	34.56 ± 0.15	73.99 ± 0.10	35.58 ± 0.08
2	65.32 ± 0.73	45.79 ± 0.74	67.28 ± 0.47	39.74 ± 0.13	64.84 ± 0.46	40.05 ± 0.17
3	58.88 ± 0.28	47.51 ± 0.36	60.98 ± 0.26	43.97 ± 0.66	57.86 ± 0.22	43.94 ± 0.13
4	54.87 ± 0.62	$\mathbf{47.64 \pm 0.58}$	55.76 ± 0.30	$\mathbf{46.15 \pm 0.15}$	52.16 ± 0.21	$\mathbf{44.90 \pm 0.20}$
5	50.20 ± 1.75	46.41 ± 0.98	50.96 ± 0.39	46.02 ± 0.27	47.46 ± 0.21	44.39 ± 0.16

TABLE III
ADVERSARIAL ATTACK SUCCESS RATES.

$n^\star$	WB-S	WB-M	WB-L
0	71.33	90.67	94.33
1	1.27 ± 0.09	0.33 ± 0.12	0.03 ± 0.01
2	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.00 \pm 0.00}$
3	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
4	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
5	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00

Evaluation metrics. The detection method is evaluated using a confusion matrix as well as the area under the receiver operating characteristic curve (AUROC) score. In the confusion matrix, a positive or negative outcome refers to an input being classified as adversarial or benign, respectively.

C. Hardware and Runtime

All experiments are performed on an NVIDIA RTX 3060 GPU with AMD Ryzen 7 5800H @ 3.20 GHz and 16 GB RAM. The runtime for the experiment can be seen in Table I. The reported times in Table I show that the proposed method might be used with constrained computational resources.

D. Results

In Tables II, III and IV, the 95% confidence intervals are reported over 10 runs, and $n^\star = 0$ indicates that the **Purifier** has not been applied.

Adversarial Purification. Comparing Clean-(S,M,L) with WB-(S,M,L) in Table II (row $n^\star = 0$) shows that the WB attack decreases relative ASR performance on clean speech by at least 71% for all sentence lengths. ASR performance on adversarial examples (or robustness accuracy) is the highest when $n^\star = 4$ for WB-(S,M,L). As seen in Table III, we do not achieve 100% adversarial attack success rate as in [16], since we did not account for the theoretical reachability of the targets. Table III further shows that WB attacks are completely defended against when $n^\star \geq 2$ for all sentence lengths.

Adversarial detection. Based on the grid search, the optimal threshold was found to be $\Omega = 0.57$ with $n^\star = 2$ for all sentence lengths. Using these hyperparameters, the accuracy is at least 87%, the true positive rate is at least 94%, and the AUROC score is at least 0.89 as shown in Table IV. Choosing $0.55 \leq \Omega \leq 0.59$ yields accuracies and AUROC scores varying approximately ± 1.00 pp, indicating that the detection algorithm is resilient to the threshold change.

IV. DISCUSSION

Tables II and III show that the **Purifier** can be used to purify adversarial examples and completely defend against WB attacks on sentences. However, the clean signals get misclassified more often when $n^\star$ increases, due to distortion.

Table IV shows that the proposed novel purification-based adversarial detection method achieves high accuracies and true positive rates, which is desirable as we aim to ensure no adversarial examples go undetected.

Methods relying on training a dedicated detection classifier, e.g., [24], achieve high detection accuracies for seen attacks, while their performance degrades dramatically for unseen examples. Our training-free approach should not be limited to specific adversarial attacks. By comparing our approach to the training-free methods in [12], [27], we achieve similar AUROC scores and accuracies with a much larger dataset.

V. CONCLUSION

In this paper, we leveraged a pre-trained DM to defend against adversarial attacks. Comprehensive experiments indicate that increasing the number of forward diffusion steps in the diffusion process improves ASR performance on adversarial examples at the cost of clean speech ASR performance. Two forward diffusion steps ensure an adversarial attack success rate of 0.00%. Finally, we introduced a novel approach utilising pre-trained DMs for detecting unknown adversarial attacks on sentences. Experiments have shown its effectiveness, achieving high AUROC scores, true positive rates, and accuracies.

TABLE IV
CLASSIFICATION SCORES FROM THE DETECTION EXPERIMENT GIVEN $\Omega = 0.57$ AND $n^\star = 2$ FOR THE WB ATTACKS. TN STANDS FOR TRUE NEGATIVE, FN FOR FALSE NEGATIVE, FP FOR FALSE POSITIVE, AND TP FOR TRUE POSITIVE.

Length		S		M		L	
TN	FP	0.76 ± 0.00	0.24 ± 0.00	$\mathbf{0.81} \pm 0.00$	$\mathbf{0.19} \pm 0.00$	0.77 ± 0.00	0.23 ± 0.00
FN	TP	$\mathbf{0.02} \pm 0.00$	$\mathbf{0.98} \pm 0.00$	0.06 ± 0.00	0.94 ± 0.00	0.04 ± 0.00	0.96 ± 0.00
Accuracy		0.87 ± 0.00		$\mathbf{0.88} \pm 0.00$		0.87 ± 0.00	
AUROC		$\mathbf{0.92} \pm 0.00$		0.90 ± 0.00		0.89 ± 0.00	

REFERENCES

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *2nd International Conference on Learning Representations*, 2014.

[2] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations*, 2015.

[3] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *2016 IEEE European symposium on security and privacy (EuroS&P)*. IEEE, 2016, pp. 372–387.

[4] M. M. Cisse, Y. Adi, N. Neverova, and J. Keshet, "Houdini: Fooling deep structured visual and speech recognition models with adversarial examples," *Advances in neural information processing systems*, vol. 30, 2017.

[5] X. Yuan, Y. Chen, Y. Zhao, Y. Long, X. Liu, K. Chen, S. Zhang, H. Huang, X. Wang, and C. A. Gunter, "{CommanderSong}: a systematic approach for practical adversarial voice recognition," in *27th USENIX security symposium (USENIX security 18)*, 2018, pp. 49–64.

[6] Y. Qin, N. Carlini, G. Cottrell, I. Goodfellow, and C. Raffel, "Imperceptible, robust, and targeted adversarial examples for automatic speech recognition," in *International conference on machine learning*. PMLR, 2019, pp. 5231–5240.

[7] R. Taori, A. Kamsetty, B. Chu, and N. Vemuri, "Targeted adversarial examples for black box audio systems," in *2019 IEEE security and privacy workshops (SPW)*. IEEE, 2019, pp. 15–20.

[8] T. Du, S. Ji, J. Li, Q. Gu, T. Wang, and R. Beyah, "Sirenattack: Generating adversarial audio for end-to-end acoustic systems," in *Proceedings of the 15th ACM Asia conference on computer and communications security*, 2020, pp. 357–369.

[9] H. Abdullah, M. S. Rahman, W. Garcia, K. Warren, A. S. Yadav, T. Shrimpton, and P. Traynor, "Hear" no evil", see" kenansville": Efficient and transferable black-box attacks on speech recognition and voice identification systems," in *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2021, pp. 712–729.

[10] G. Chen, S. Chenb, L. Fan, X. Du, Z. Zhao, F. Song, and Y. Liu, "Who is real bob? adversarial attacks on speaker recognition systems," in *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2021, pp. 694–711.

[11] C. Herff and T. Schultz, "Automatic speech recognition from neural signals: a focused review," *Frontiers in neuroscience*, vol. 10, p. 429, 2016.

[12] Z. Yang, B. Li, P.-Y. Chen, and D. Song, "Characterizing audio adversarial examples using temporal dependency," *International Conference on Learning Representations*, 2019.

[13] K. Rajaratnam, B. Alshemali, and J. Kalita, "Speech coding and audio preprocessing for mitigating and detecting audio adversarial examples on automatic speech recognition," 2018.

[14] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *2016 IEEE symposium on security and privacy (SP)*. IEEE, 2016, pp. 582–597.

[15] G. Chen, Z. Zhao, F. Song, S. Chen, L. Fan, F. Wang, and J. Wang, "Towards understanding and mitigating audio adversarial examples for speaker recognition," *IEEE Transactions on Dependable and Secure Computing*, vol. 20, no. 5, pp. 3970–3987, 2022.

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

[17] ——, "Towards evaluating the robustness of neural networks," in *2017 IEEE symposium on security and privacy (sp)*. IEEE, 2017, pp. 39–57.

[18] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *Proceedings of the International Conference on Representation Learning*, 2017.

[19] S. Wu, J. Wang, W. Ping, W. Nie, and C. Xiao, "Defending against adversarial audio via diffusion model," *The Eleventh International Conference on Learning Representations*, 2023.

[20] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International conference on machine learning*. PMLR, 2015, pp. 2256–2265.

[21] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[22] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8162–8171.

[23] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "Diffwave: A versatile diffusion model for audio synthesis," *ICLR 2021 (oral)*, 2021.

[24] S. Samizade, Z.-H. Tan, C. Shen, and X. Guan, "Adversarial example detection by classification for deep speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3102–3106.

[25] C. H. Nielsen and Z.-H. Tan, "Leveraging domain features for detecting adversarial attacks against deep speech recognition in noise," *IEEE Open Journal of Signal Processing*, vol. 4, pp. 179–187, 2023.

[26] S. Hussain, P. Neekhara, S. Dubnov, J. McAuley, and F. Koushanfar, "WaveGuard: Understanding and mitigating audio adversarial examples," in *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, Aug. 2021, pp. 2273–2290.

[27] H. Kwon and S.-H. Nam, "Audio adversarial detection through classification score on speech recognition systems," *Computers & Security*, vol. 126, p. 103061, 2023.

[28] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2020, pp. 4211–4215.

[29] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine learning*, 2006, pp. 369–376.

[30] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations*, 2014.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[32] Z.-H. Tan, A. kr. Sarkar, and N. Dehak, "rVAD: An unsupervised segment-based robust voice activity detection method," *Computer speech & language*, vol. 59, pp. 1–21, 2020, https://github.com/zhenghuatan/rVAD.

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