ADVERSARIAL ATTACKS AGAINST AUDIO SURVEILLANCE SYSTEMS

Stavros Ntalampiras

University of Milan, Department of Computer Science, via Celoria 18, 20133, Milan, Italy

ABSTRACT

The recent rise of adversarial machine learning highlights the vulnerabilities of various systems relevant in a wide range of application domains. This paper focuses on the important domain of automatic space surveillance based on the acoustic modality. After setting up a state of the art solution using log-Mel spectrogram modeled by a convolutional neural network, we systematically investigate the following four types of adversarial attacks: *a*) Fast Gradient Sign, *b*) Projected Gradient Descent, *c*) Jacobian Saliency Map, and *d*) Carlini & Wagner ℓ_∞. Experimental scenarios aiming at inducing false positives or negatives are considered, while attacks' efficiency are thoroughly examined. It is shown that several attack types are able to reach high success rate levels by injecting relatively small perturbations on the original audio signals. This underlines the need of suitable and effective defense strategies, which will boost reliability in machine learning based solution.

Index Terms— adversarial machine learning, audio signal processing, convolutional neural network, acoustic surveillance, urban environment.

1. INTRODUCTION

Even though machine learning algorithms are being developed and deployed into commercial products on a daily basis, recent research has shown that they could be vulnerable to malicious attempts, so-called *adversarial attacks* [1–3]. Following their overwhelming success in a wide variety of fields including audio signal processing [4], deep learning based approaches are the ones primarily targeted. The typical attack scenario includes applying small perturbations onto the testing data (creating so-called *adversarial examples*), so that the existing model is led to misclassifications. Such purposefully designed challenges pose significant threats with respect to the implementation and deployment of ML-based solutions. Combined with the limited interpretability of many current deep nets [5], adversarial attacks contribute to the uncertainty surrounding AI when it comes to real-world critical applications (e.g. healthcare [6], critical infrastructures [7], to mention but a few) and especially its trustworthiness [8].

As a rule, the goal of such attacks is to alter the prediction made by an ML-based classification system by modifying the input's content as less as possible [9]. Attacks are broadly categorized as *untargeted* and *targeted* [10]: in case of untargeted attacks, the attacker aims at causing a misclassification without any constraints on the identity of the predicted class. On the contrary, in case of targeted attacks, the attacker aims that a specific class is predicted by the existing model. A further differentiation concerns the available knowledge regarding the ML-based classifier: in case the attacker is fully aware of the model (architecture, node weights, etc.) the attack falls in the *white-box* category. In case the attacker does not have any such information available, the attack belongs to the *black-box* category. There, a so-called *surrogate* model is employed to design the attacks, which is then deployed based on the transferability of adversarial examples [11]. Lastly, given the information available to the attacker (architecture, part of training data, algorithm, etc.), *grey-box* attacks can be derived as well.

This work is concentrated on adversarial attacks performed against audio surveillance systems, where the classes are representative of several typical and atypical situations. Unlike the image processing domain where there is already a substantial amount of research on adversarial machine learning, similar attempts are rather limited on audio signals and even more so on non-speech audio [12].

Focusing on speech signals, the authors of [13] present an attack methodology able to be carried out in the physical world meaning that the environment may potentially be characterized by reverberation and/or noise. In this direction, [14] presents a physically realizable adversarial attack based on a psychoacoustic property-based loss function. Attacks are generated considering room impulse responses of different environments. Interestingly, attacks against voice assistants can also be found in the literature [15]. There, the goal was to deactivate the assistant as long as the adversarial example is present. Another class of attacks to generate adversarial audio is presented in [16].

Unfortunately, the amount of works focused on non-speech audio is rather limited. Subramaninan et al. [17] present white-box attack algorithms along with baseline defenses. They conclude that adversarial attacks are not robust to simple input transformations with white noise being the most efficient defense strategy. With respect to music signals, the authors of [18] demonstrate how adversarial attacks commonly-used in images are able to fool a 2D convolutional neural network trained for music genre classification. At the same time, audio reconstructed from perturbed spectrograms is not perceived as dissimilar to the legitimate one by human listeners. Finally, Esmaeilpour et a. [19] explain an SVM-based environmental sound classification system fed on discrete wavelet transform representations offering a satisfactory trade-off between accuracy and resilience.

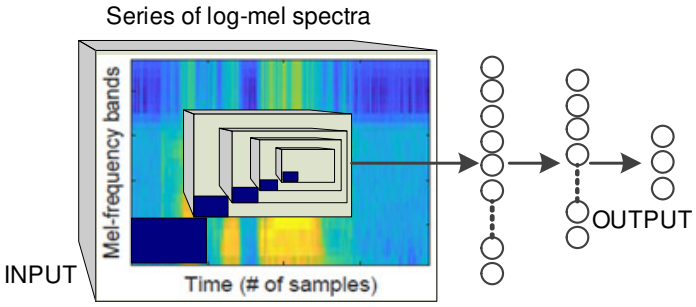

Fig. 1. The CNN architecture for acoustic surveillance of hazardous situations.

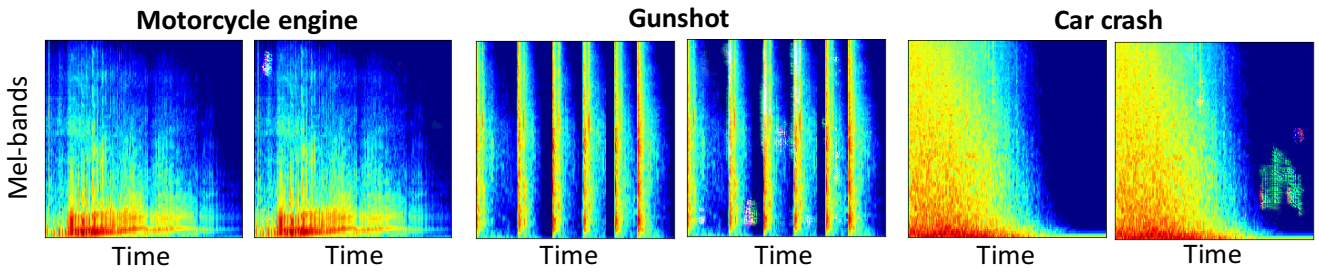

Fig. 2. Original and attacked log-Mel spectrograms. First pair: motorcycle engine and successful attacked version based on *JSM* with $\epsilon = 0.3$. Second pair: gunshot and failed attacked version based on *JSM* with $\epsilon = 0.5$. Third pair: car crash and failed attacked version based on *JSM* with $\epsilon = 0.5$. We see that the attack alters a small number of the log-Mel spectrogram components; these are visible in upper-left part of the motorcycle engine sound, lower-central part of the gunshot sound, and lower right part of the car crash sound. We observe the tendency to add time-frequency elements than distort existing ones, while lower frequency parts are preferred.

This work examines systematically the efficacy of the most prominent adversarial attacks applicable onto acoustic surveillance systems. These hold a central role in security applications making a systematic study necessary in order to understand and assess these potential vulnerabilities. It represents a realistic application scenario of critical relevance with potentially direct consequences on human lives and/or property loss/damage. We assume an urban soundscene where typical events include vehicles passing by, weather events, etc., and the atypical situations are car crashes and gunshots. Seeing the problem from the attacker's perspective, we introduce attacks aiming at misclassifying typical events as atypical and vice versa. Following such requirements, the attack has to be targeted without any knowledge regarding the model used to analyze the urban soundscene. At the same time, it is not reasonable to assume that the attacker has access to the legitimate training data, and thus is unable to carry out injection type of attacks [20]. To this end, we included the following four attack types *a*) Fast Gradient Sign, *b*) Projected Gradient Descent, *c*) Jacobian Saliency Map, and *d*) Carlini & Wagner ℓ_∞. We carried out extensive experiments to thoroughly assess the success of these attacks in the present scenario, while quantifying the applied perturbations and confidence levels.

2. PROBLEM FORMULATION

Let us define a classification task $\mathcal{T}$ associated with a known set of audio classes $\mathcal{C} = \{C_1, \ldots, C_m\}$, where m is the number of classes and a classification model $\mathcal{M}$. In the proposed surveillance application, part of $\mathcal{C}$ is associated with typical events ($\mathcal{C}^T$) while the rest $\mathcal{C}^A$ with atypical, potentially hazardous situations ($\mathcal{C}^T \cup \mathcal{C}^A = \mathcal{C}$). Following the recent success of spectrogram-based audio analysis, without loss of generality, we assume inputs x of image form, i.e. $x \in \mathbb{R}^{d_1,d_2,d_3}$, where d_1 is the width, d_2 the height, and d_3 the number of color channels of the image. The overall goal of the surveillance framework is to classify the incoming sounds while minimizing false positive and negative rates.

In such a context, the attack includes an ideally imperceptible perturbation $\psi(x, y)$, where x is the input to be perturbed and y the target class. Contrary to the acoustic surveillance framework, the goal of an attacker would include targeted attacks aiming at "hiding" the existence of abnormal situations in $\mathcal{C}^A$, i.e. increasing the false negative rate, and/or make $\mathcal{M}$ falsely classify an event as abnormal even though it is normal, thus raising a false alarm.

This work investigates the efficacy of several perturbation functions ψ in the audio surveillance domain and assess their efficacy using a state of the art audio classification method.

3. THE AUDIO CLASSIFICATION SYSTEM

Motivated by the related literature [4, 21], the classification algorithm first extracts log-Mel spectrograms, the distribution of which is learned by a Convolutional Neural Network facilitating both modeling and inferring processes. The following subsections explain briefly the feature extraction and pattern recognition stages, along with the employed dataset and achieved results.

3.1. Mel-spectrogram

We employed 40 equal-width log-energies with an overlapping based on the Mel filterbank. The standard extraction method is followed including the computation of short time Fourier transform. Mel-spectrograms have been proven to emphasize components which play an important role to human perception [4].

3.2. The Convolutional neural network

The present CNN is composed of four convolutional layers followed by max-pooling layers (see Fig. 1). The network further includes a flattening, a dropout and three densely connected layers, where the regression process is carried out. In order to avoid overfitting, a dropout layer was considered randomly removing 50% of the present hidden units. The specific process scales down the number of parameters to estimate during learning, while discarding insignificant relationships.

The present topology includes 3 convolutional layers using the standard ReLU activation function with a total number of parameters equal to 2,197,667. The kernel size is 3×3 with stride being equal to

Table 1. The confusion matrix (in %) achieved by audio classification system approach. The average recognition rate is 99.1%. The highest rates per species are emboldened.

Presented \ **Responded**	*back*	*gunshot*	*car crash*
back	**100**	-	-
gunshot	-	**99.2**	0.8
car crash	1.1	0.7	**98.2**

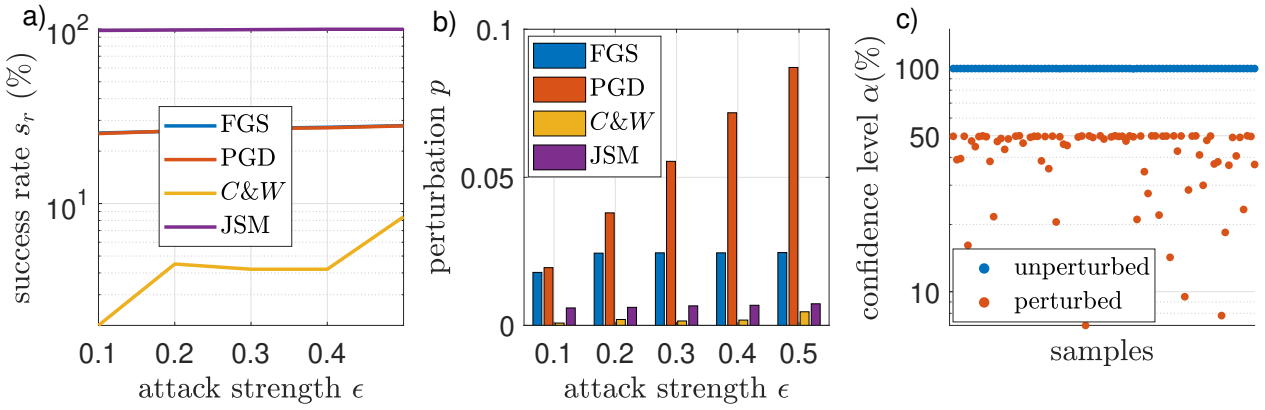

Fig. 3. This scenario perturbs typical urban background events targeting to have them misclassified as atypical, i.e. inducing false alarms. a) Success rate (%) perturbation vs. attack strength ϵ for the considered attack types. b) Normalized perturbation vs. attack strength ϵ for the considered attack types. c) Confidence level α (%) for original typical and attacked atypical samples using *JSM* with $\epsilon = 0.3$.

2, while the number of filters varies between 32 and 64. The number of output nodes represents the considered classes and is equal to 3 as shown in Fig. 1. It should be mentioned that the topology is optimized in terms of hyperparameters based on grid search including early stopping.

3.3. Dataset and Classification Results

The dataset contains audio acquired from professional sound effects collections (BBC Sound Effects Library, Sound Ideas Series 6000, Sound Ideas: the art of Foley, and Best Service Studio Box Sound Effects). These kinds of collections comprise an enormous source of high quality recordings used by the movie industry [22]. The considered classes are *a*) background urban environmental sounds (typical), such as cars/motorcycles passing by, rain, wind, crowd, etc. *b*) gunshot (atypical), and *c*) car crash (atypical). The total durations are 9968.2s, 2290.75s, and 1975.2s for each class respectively.The sampling rate is 16 kHz with 16 bit quantization.

The achieved confusion matrix following a ten-fold cross validation experimental protocol tabulated in Table 1, demonstrates excellent classification performance reaching an overall percentage of 99.1%. This confirms the efficiency of the designed classification solution for the task-at-hand.

4. THE CONSIDERED ATTACKS

The attack types considered in this work are outlined in this section. We aimed at assessing the success of each one against the audio surveillance framework, thus we selected every attack available in the literature which can be applied in a targeted manner, while being agnostic with respect to the classification model (see section 2).

a) Fast Gradient Sign (*FGS*) [23] can be used both for targeted and untargeted attacks as it tries to control the ℓ_1, ℓ_2 or ℓ_∞ norm of the change injected by the adversary. In the specific case, $\psi(x, y) = -\epsilon \times \text{sign}\nabla_x(\mathcal{L}(x, y))$, where $\epsilon > 0$ denotes the attack strength, $\mathcal{L}$ the loss function, and y the target class as defined in section 2. The aim of this attack is to apply alterations on x so that the classifier's loss is minimized when it's prediction is y. Several values of ϵ are considered in combination with log-Mel spectrogram's perturbation.

b) Projected Gradient Descent (*PGD*) [24] is an iterative version of FGS, where the attack is executed several times in an iterative manner. The overall attack strength is again ϵ. There is an additional parameter called ϵ_{step} defining each iteration's step size. During each iteration, the result of the attack is projected onto the ϵ-norm sphere, the center of which is the original input x.

c) Jacobian Saliency Map (*JSM*) [25]: Unlike FGS, this attack tries to control the ℓ_0 norm of the change injected by the adversary. JSM alters a predermined number of x's components (limited by an upper bound δ) while constructing the ψ. This process is carried out iteratively until either a) δ is reached, or b) the targeted erroneous prediction is accomplished.

d) Carlini & Wagner ℓ_∞ (*C&W*) [26]: this attack focuses on the minimization of ℓ_∞ norm of adversarial samples. Overall, it searches for the optimal trade-off between accomplishing the targeted misclassification, while preserving the perturbation $\psi(x, y)$ as small as possible. It wishes to discover a perturbed signal $\psi(x, y)$ with $\ell(\psi) = 0$ (ℓ being the same with the ℓ_2 version). At the same time, the condition $||x - \psi||_\infty \leq \epsilon$, with $\epsilon > 0$ has to met. Essentially, ϵ is the amount of permitted perturbation.

5. EXPERIMENTS

This section describes *a*) the employed dataset, *b*) the parameterization of the included modules, and *c*) the experimental protocol and results for the considered attack scenarios and human auditory experiment. The code to reproduce the experiments is available at `https://sites.google.com/site/stavrosntalampiras/software`.

5.1. Features and model parameterization

DC offset was removed from all files in the dataset. Following MPEG-7 standard recommendations, feature extraction frame is 30 ms with 10 ms step. Hamming windowing is applied, while the FFT size was 1024. In addition, standard normalization techniques, i.e. mean removal and variance scaling, were used. Attack strength ϵ was taken from the set $\{0.1, 0.2, 0.3, 0.4, 0.5\}$ while we employed the implementation provided in [10]. Lastly, CNN training process is bounded by 1000 epochs at a learning rate of 0.0001.

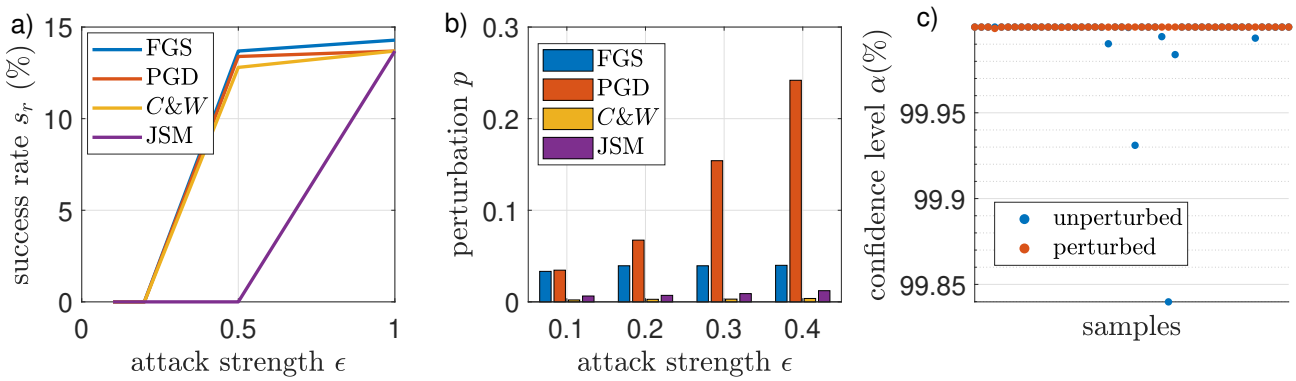

Fig. 4. This scenario attacks atypical sound events targeting to be misclassified as typical, i.e. increasing false negatives. a) Success rate (%) perturbation vs. attack strength ϵ for the considered attack types. b) normalized perturbation vs. attack strength ϵ for the considered attack types. c) Confidence level α (%) achieved for original atypical and perturbed typical samples using *FGS* with $\epsilon = 0.3$.

5.2. Experimental protocol and results

Following the motivation described in sections 1 and 2, we focus on two attack application scenarios: 1. *AAS1*: targeted attacks forcing $\mathcal{M}$ to falsely classify an event as abnormal even though it is normal, i.e. increasing false positive rate, and 2. *AAS2*: targeted attacks aiming at "hiding" the existence of abnormal situations in $\mathcal{C}^A$, i.e. increasing the false negative rate. We followed the ten-fold cross-validation division as in section 3 where the testing data of each fold was perturbed by the attack types described is section 4. Thus, all samples available in the dataset were employed for every attack and the results are averaged. The performance of each attack type is measure by means of three figures of merit: 1. success rate s_r which counts the times an attack was successful, i.e. the target class was predicted by the model (percentage), and 2. perturbation p which is the absolute difference between the original x and the adversarial sample with alteration ψ. 3. classification confidence $\alpha = p(y|\mathcal{M})/\sum p(*|\mathcal{M})$, which is the probability of the predicted class divided by the sum of all probabilities.

5.3. Attack application scenario 1

The results obtained w.r.t *AAS1* are demonstrated in Fig. 3. Fig. 3a shows the success rate vs. various values of ϵ with respect to every attack type. We observe that s_r exhibits a constant behavior as ϵ increases from 0.1 to 0.5. In other words, it seems that the attack strength does not have a significant influence on its success rate, even if perturbation increases. The attack with the lowest s_r is *C&W*, medium (and similar) s_r by *FGS* and PGD, while the attack with the highest s_r is *JSM*. Interestingly, *JSM* achieved nearly perfect performance when applied with the lowest strength ϵ and it was excellent with strength equal to 0.5. At the same time, Fig. 3b shows how p alters for the same values of ϵ. In general, p increases with attack strength exhibiting a quasi-linear relationship. The attack introducing the largest perturbation is *PSG*, while *C&W* the smallest. Interestingly, the attack with the highest s_r exhibits relatively small perturbations across all strengths. More intense attacks inducing more serious perturbations were not considered, since *JSM* was able to reach perfect success rate with $\epsilon = 0.5$. Finally, Fig. 3c demonstrates the classification confidence associated to the original, i.e. typical background class. There, we see that α values associated with original unperturbed samples are higher than the ones corresponding to perturbed samples. This could comprise the basis for a defense strategy against *AAS1* attacks. Fig. 2 visualizes such an attack: the first pair of images represents the log-Mel spectrogram of a motorcycle engine sound and its attacked version using *JSM* with $\epsilon = 0.3$.

5.4. Attack application scenario 2

The results obtained w.r.t *AAS2* are demonstrated in Fig. 4. Fig. 4a shows the success rate vs. various values of ϵ with respect to every attack type. We observe that s_r surges as ϵ increases from 0.1 to 1. Higher attack strengths were not considered due to the excessive perturbations they induced. *JSM* does not reach the levels observed in $AAS1$. Here, *FGS* provides the highest s_r which is approximately 15% with $\epsilon = 1$ injecting a perturbation $p \simeq 0.05$. Fig. 4b demonstrated the corresponding perturbations, which increase with ϵ. Similar to $AAS1$, *PSG* introduces the largest perturbation and *C&W* the smallest. Overall, s_r's achieved in $AAS2$ are lower than those in $AAS1$. Interestingly, the confidence levels shown in Fig. 4c are similar for perturbed and unperturbed samples, thus a different defense strategy with respect to *AAS1* should be applied here. In this case, confidence α is associated to the original class, i.e. atypical (gunshot/car crash). In general, success rates in AAS1 are higher than AAS2 which might be associated with the inherent difficulty in smoothing clipping sound events. Fig. 2 visualizes such attacks: the second and third pairs of images represents the logMel spectrogram of a gunshot and a car crash sound along with their attacked versions using *JSM* with $\epsilon = 0.5$.

6. CONCLUSIONS

This work demonstrated the applicability of adversarial attacks on the critical domain of acoustic surveillance. After a carefully designed experimental process considering diverse aspects of the specific problem, it highlighted that such systems are indeed vulnerable, especially when the attacker wishes to inject false alarms.

In the future, we are going to assess the performance of adversarial attacks under potential reverberation effects [27]. Moreover, we wish to explore a wide range of defense strategies, e.g. based on psychoacoustic principles [28], audio steganography [29], exploiting temporal dependencies [30], etc. Another relevant research direction is to examine the performance of such attacks in speech signals. Finally, we are going to investigate the transferability of adversarial attacks across models with diverse architectures, parameters, and training datasets.

7. REFERENCES

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *International Conference on Learning Representations*, 2014. [Online]. Available: http://arxiv.org/abs/1312.6199

[2] J. Han, Z. Zhang, N. Cummins, and B. W. Schuller, "Adversarial training in affective computing and sentiment analysis: Recent advances and perspectives," *CoRR*, vol. abs/1809.08927, 2018.

[3] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6572

[4] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S. Chang, and T. Sainath, "Deep learning for audio signal processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 206–219, 2019.

[5] C. Wu, M. J. F. Gales, A. Ragni, P. Karanasou, and K. C. Sim, "Improving interpretability and regularization in deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 256–265, 2018.

[6] P. Tyrväinen, M. Silvennoinen, K. Talvitie-Lamberg, A. Ala-Kitula, and R. Kuoremäki, "Identifying opportunities for ai applications in healthcare — renewing the national healthcare and social services," in *2018 IEEE SeGAH*, 2018, pp. 1–7.

[7] S. Ntalampiras, "Fault identification in distributed sensor networks based on universal probabilistic modeling," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 9, pp. 1939–1949, 2015.

[8] S. A. Seshia, S. Jha, and T. Dreossi, "Semantic adversarial deep learning," *IEEE Design Test*, pp. 1–1, 2020.

[9] E. Tabassi, K. J. Burns, M. Hadjimichael, A. D. Molina-Markham, and J. T. Sexton, "A taxonomy and terminology of adversarial machine learning," Oct. 2019. [Online]. Available: https://doi.org/10.6028/nist.ir.8269-draft

[10] M.-I. Nicolae, M. Sinn, M. N. Tran, B. Buesser, A. Rawat, M. Wistuba, V. Zantedeschi, N. Baracaldo, B. Chen, H. Ludwig, I. Molloy, and B. Edwards, "Adversarial robustness toolbox v1.2.0," *CoRR*, vol. 1807.01069, 2018.

[11] F. Tramèr, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "The space of transferable adversarial examples," *arXiv*, 2017. [Online]. Available: https://arxiv.org/abs/1704.03453

[12] H. Kwon, Y. Kim, H. Yoon, and D. Choi, "Selective audio adversarial example in evasion attack on speech recognition system," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 526–538, 2020.

[13] H. Yakura and J. Sakuma, "Robust audio adversarial example for a physical attack," *CoRR*, vol. abs/1810.11793, 2018. [Online]. Available: http://arxiv.org/abs/1810.11793

[14] J. Szurley and J. Z. Kolter, "Perceptual based adversarial audio attacks," *ArXiv*, vol. abs/1906.06355, 2019.

[15] B. L. junchenl, B. C. bingqinc, and Z. Z. zhuoran, "Adversarial music : Real world audio adversary against wake-word detection systems," 2019.

[16] T. Du, S. Ji, J. Li, Q. Gu, T. Wang, and R. Beyah, "Sirenattack: Generating adversarial audio for end-to-end acoustic systems," *Proceedings of the 15th ACM Asia Conference on Computer and Communications Security*, 2020.

[17] V. Subramanian, E. Benetos, N. Xu, S. McDonald, and M. B. Sandler, "Adversarial attacks in sound event classification," *CoRR*, vol. abs/1907.02477, 2019.

[18] K. M. Koerich, M. Esmailpour, S. Abdoli, A. S. Britto, and A. L. Koerich, "Cross-representation transferability of adversarial perturbations: From spectrograms to audio waveforms," *ArXiv*, vol. abs/1910.10106, 2019.

[19] M. Esmaeilpour, P. Cardinal, and A. L. Koerich, "A robust approach for securing audio classification against adversarial attacks," *IEEE Trans. on Information Forensics and Security*, vol. 15, pp. 2147–2159, 2020.

[20] Z. He, A. S. Rakin, and D. Fan, "Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack," in *2019 IEEE/CVF CVPR*, 2019, pp. 588–597.

[21] S. Ntalampiras, "Moving vehicle classification using wireless acoustic sensor networks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 129–138, April 2018.

[22] S. Ntalampiras, D. Arsić, M. Hofmann, M. Andersson, and T. Ganchev, "PROMETHEUS: heterogeneous sensor database in support of research on human behavioral patterns in unrestricted environments," *Signal, Image and Video Processing*, vol. 8, no. 7, pp. 1211–1231, Jun. 2012.

[23] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, cite arxiv:1412.6572. [Online]. Available: http://arxiv.org/abs/1412.6572

[24] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[25] 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 EuroSP*, 2016, pp. 372–387.

[26] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 39–57.

[27] S. Ntalampiras, "Generalized sound recognition in reverberant environments," *J. Audio Eng. Soc*, vol. 67, no. 10, pp. 772–781, 2019.

[28] N. Das, M. Shanbhogue, S.-T. Chen, L. Chen, M. E. Kounavis, and D. H. Chau, "ADAGIO: Interactive experimentation with adversarial attack and defense for audio," in *Machine Learning and Knowledge Discovery in Databases*. Springer International Publishing, 2019, pp. 677–681. [Online]. Available: https://doi.org/10.1007/978-3-030-10997-4_50

[29] J. Wu, B. Chen, W. Luo, and Y. Fang, "Audio steganography based on iterative adversarial attacks against convolutional neural networks," *IEEE Trans. on Information Forensics and Security*, vol. 15, pp. 2282–2294, 2020.

[30] Z. Yang, B. Li, P. Chen, and D. Song, "Characterizing audio adversarial examples using temporal dependency," *CoRR*, vol. abs/1809.10875, 2018. [Online]. Available: http://arxiv.org/abs/1809.10875