On the Robustness of State-of-the-Art Transformers for Sound Event Classification against Black Box Adversarial Attacks

Christos Nikou, Vasileios Theiou, Stefanos Vlachos, Christos Sgouropoulos,
Dimitris Sgouropoulos, Theodoros Giannakopoulos

Multimedia Analysis Group of the Computational Intelligence Laboratory (MagCIL)
Institute of Informatics and Telecommunications, NCSR "DEMOKRITOS"

***Abstract*—Deep learning models are now ubiquitous in many applications, yet they remain vulnerable to adversarial examples. The robustness of transformer-based models in the audio domain, however, has not been thoroughly investigated. To address this, in this paper, we evaluate the robustness of three state-of-the-art pretrained transformer-based architectures for sound event classification. We propose a method to generate adversarial examples with a fixed signal-to-noise ratio in a black-box setting, utilizing an evolutionary algorithm. This approach enables a refined assessment of model robustness against varying levels of imperceptibility. To ensure statistical significance and variability, we conduct extensive experiments using two benchmark datasets, reporting success rates from ~10% to ~95% depending on SNR. Our findings reveal significant vulnerabilities in current state-of-the-art transformer models, demonstrating that, similar to the image domain, model performance may not correlate with robustness. These results underscore the need to re-evaluate both the performance and robustness of such models. We publicly release our code and adversarial examples in https://magcil.github.io/audio-adversarial-attacks/, showcasing the correlation between SNR and imperceptibility.**

***Index Terms*—Deep Learning, Sound Event Classification, Robustness, Adversarial Attacks.**

I. Introduction

Advancements in deep learning have significantly improved the performance of sound event classification systems. State-of-the-art (SOTA) deep learning models now achieve remarkable accuracy across diverse audio datasets. However, their vulnerability to adversarial examples—carefully crafted inputs designed to mislead models—has emerged as a critical concern. These attacks pose a serious threat to system security, enabling attackers to manipulate decisions by introducing perturbations that are imperceptible to human listeners. Such vulnerabilities highlight the need to understand why these models are susceptible to such inputs, and re-evaluate the performance and robustness before deploying them in real-world applications.

Szegedy et al. [1] initiated the field of adversarial attacks by demonstrating the existence of adversarial examples in the image domain. This work led to the development of a series of methods for crafting adversarial examples. These methods can be broadly classified into white-box approaches [2]–[4], where the adversary has full access to the model, and black-box approaches [5]–[7], where the target model is unknown. Based on these studies, researchers have extended adversarial techniques to the audio domain, primarily targeting Automatic Speech Recognition (ASR) systems [8]–[10]. However, the exploration of adversarial attacks in non-speech-related audio tasks remains limited. In the context of sound event classification, recent works [11]–[15] have focused on generating adversarial examples that can deceive machine learning models while remaining imperceptible to human listeners. Although these studies demonstrate the vulnerability of sound event detection systems to adversarial attacks, they do not consider modern transformer-based architectures, which represent the current SOTA in many audio classification tasks [16]–[18]. To address these limitations, in this paper, we evaluate the robustness of three SOTA transformer-based architectures for sound event classification. In detail, our contributions are summarized as follows:

- To the best of our knowledge, this is the first work to evaluate the robustness of SOTA transformer-based models for sound event classification against adversarial attacks. Through our experiments, we demonstrate that model performance is uncorrelated with robustness. This phenomenon, previously observed in the image domain [19], is now also verified in the audio domain.
- We utilize an evolutionary algorithm to generate adversarial examples in a black-box setting, a realistic and practical scenario. Although this approach has been employed in prior works [10], [13], we refine it to generate adversarial examples with a fixed Signal-to-Noise Ratio (SNR). Additionally, we introduce an initialization strategy that mimics the auditory masking effect. These modifications enable a comprehensive evaluation of model robustness against adversarial examples at varying levels of imperceptibility.
- We perform large-scale experiments on two benchmark datasets to ensure statistical significance, enhancing the reliability of our findings.

The structure of this paper is as follows: In Section II, we

present the adopted methodology for generating adversarial examples with fixed SNR. Section III details the experimental setup, including the datasets used, preprocessing steps, the models under attack, and the implementation specifics. In Section IV, we present and analyze the experimental results. Finally, Section V concludes the paper by summarizing the key findings and discussing potential directions for future research.

II. Methodology

Notation: Let $f : \mathbb{R}^D \to [0,1]^K$ be a neural network representing the sound event detector. The detector takes as input an audio waveform $x \in \mathbb{R}^D$, and outputs a $K-$dimensional probability vector $f(x)$, where K denotes the number of predicted classes. The vector $f(x)$ corresponds to the posteriors of the neural network, and $f(x)_k$ denotes the k^{th} coordinate of $f(x)$. For an audio event $x \in \mathbb{R}^D$, the prediction of the model is given by $c(x) = \arg\max_{1\le k\le K} f(x)_k$.

A. Black-Box Audio Adversarial attacks

General Formulation: In the black-box setting, the only information that is accessible is the probability vector $f(x)$. The adversary's objective is to find a perturbation $\delta^* \in \mathbb{R}^D$ such that the event $x_{\text{adv}} = x+\delta^*$ is misclassified by the model, i.e., $c(x) \neq c(x_{\text{adv}})$. A critical constraint is that δ^* must remain imperceptible, meaning it should not be detectable by human observation. In this case the optimal solution δ^* is formulated as $\delta^* = \arg\min_{\delta\in\mathbb{R}^D} \|\delta\|$ such that $c(x+\delta) \neq c(x)$ for all $\delta \in \mathbb{R}^D$. Thus, imperceptibility can be achieved by minimizing the perturbation with respect to a chosen norm $\|\cdot\|$.

Reformulation: The intractability of the above problem due to the constraint $c(x+\delta) \neq c(x)$ leads to an alternative reformulation to approximate δ^*. To this end, to find the perturbations δ that satisfy $c(x+\delta) \neq c(x)$ we minimize the loss function

$$L(x_{\text{adv}}) = f(x_{\text{adv}})_{c(x)} - \max_{\substack{1\le k\le K \\ k\neq c(x)}} f(x_{\text{adv}})_k. \tag{1}$$

The attack is successful when $L(x_{\text{adv}}) < 0$. In this case, there will be at least one posterior $f(x_{\text{adv}})_k > f(x_{\text{adv}})_{c(x)}$, with $k \neq c(x)$, which implies $c(x_{\text{adv}}) \neq c(x)$.

B. Optimization Algorithms

To minimize the loss function L in (1) we employ a meta-heuristic optimization strategy. In detail, we utilize Particle Swarm Optimization (PSO) [20], a population-based evolutionary algorithm. Additionally, we normalize the adversarial perturbation to achieve a fixed SNR. This framework is general and can be used with any evolutionary algorithm, such as Differential Evolution (DE) [21]. In our experiments, we utilized DE as well and did not observe any difference; hence, in this manuscript we present only PSO.

Initialization: While the SNR constraint is a necessary condition for imperceptibility, it is not always sufficient. For example, if the adversarial perturbation contains high-energy components in perceptually salient regions of the original input, it may become noticeable despite satisfying the SNR constraint. To address this, we initialize the perturbations based on the input waveform, leveraging principles from the auditory masking effect—where a strong signal can hide nearby lower-amplitude components from human perception. Formally, for an audio event $x \in \mathbb{R}^D$, the i^{th} perturbation is initialized as $\delta_i(n) = W \cdot \text{sign}(x(n)) \cdot \text{rand}\,(0, |x(n)|)$, where $\text{sign}(\cdot)$ is the sign function, $\text{rand}(\alpha,\beta)$ denoted a uniformly chosen number in the interval (α,β), and $W \in (0,1)$ is a scaling factor. This enables the injection of greater noise into high-amplitude segments of the signal, where audiotory masking is more effective. Fig. 1 illustrates this phenomenon.

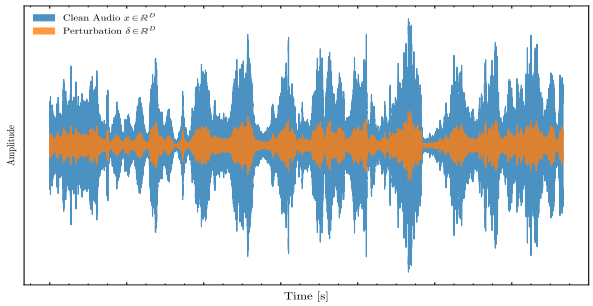

Fig. 1. Clean audio $x \in \mathbb{R}^D$, and adversarial perturbation $\delta^* \in \mathbb{R}^D$. The example $x_{\text{adv}} = x+\delta^*$ is classified as "silence", while corresponds to human speech.

SNR Constraint: In the image domain, l_{p} norms quantify the noise magnitude, but in audio, SNR is a more suitable metric to evaluate and regulate noise levels [22]. The SNR is defined as $\text{SNR (dB)} = 10\cdot\log_{10}(E_y/E_n)$, where E_y, E_n are the energies of the clean signal y_{signal}, and noise signal y_{noise}, respectively. To obtain an adversarial example of fixed SNR, we normalize the noise before adding it to the clean signal. In detail, for a given SNR value S measured in dB, we define the scaling factor $\alpha_{y,n}(S) = \sqrt{\frac{E_y}{E_n}\cdot 10^{-S/10}}$. Then, the signal $z = y_{\text{signal}} + \alpha_{y,n}(S)\cdot y_{\text{noise}}$ has SNR equal to S.

Particle Swarm Optimization algorithm initializes a swarm of particles, each representing a potential adversarial solution in the search space. Particles update their velocities and positions iteratively, influenced by their personal best, and global best positions. Let P denote the swarm size, and D the dimensionality of the audio waveform. At time step t, each particle $1 \le i \le P$ has a position $x_i \in \mathbb{R}^D$, and a velocity $v_i \in \mathbb{R}^D$. The velocity update is given by $v_i^{t+1} = wv_i^t + c_1 r_1^{t+1}(p_i^{\text{best}} - x_i^t) + c_2 r_2^{t+1}(g^{\text{best}} - x_i^t)$, where p_i^{best} is the particle's i best position, g^{best} is the global best, w is the inertia weight and c_1, c_2 control the influence of personal and global best positions. The random factors $r_1^{t+1}, r_2^{t+1} \in U(0,1)$, introduce stochasticity. The position update is $x_i^{t+1} = x_i^t + v_i^{t+1}$. Algorithm 1 presents the workflow of PSO in pseudocode.

III. Experimental Setup

A. Datasets

AudioSet [23] is a large-scale dataset of manually annotated audio events, comprising 527 distinct classes organized in a

Algorithm 1 Particle Swarm Optimization

Input: Event $x \in \mathbb{R}^D$, particles P, iterations N, hyperparameters w, c_1, c_2, scaling factor W, SNR S.
Initialize Swarm Best Fitness: $SBF \leftarrow \infty$
for each $i = 1, \ldots, P$ **do**
  Initialize perturbation $\delta_i^{(0)}$.
  Set $x_i^0 \leftarrow x + \delta_i^0, v_i^0 \leftarrow x_i^0$, and $p_i^{\text{best}} \leftarrow x_i$.
  if $L(x_i^0) < SBF$ **then** update $SBF \leftarrow L(x_i^0)$, and $\mathbf{g}^{\text{best}} \leftarrow x_i^0$.
end for
Initialize iteration index $t \leftarrow 0$.
while $SBF > 0$ or $t < N$ **do**
  for $i = 1, \ldots, P$ **do**
    Update velocity: $v_i^{t+1} \leftarrow w v_i^t + c_1 r_1 (p_i^{\text{best}} - x_i^t) + c_2 r_2 (\mathbf{g}^{\text{best}} - x_i^t)$.
    Perturbation: $\delta_i^{t+1} \leftarrow \delta_i^t + v_i^{t+1}$.
    Position: $x_i^{t+1} \leftarrow x_i^t + \alpha_{x_i^{t+1}, \delta_i^{t+1}}(S) \cdot \delta_i^{t+1}$.
    if $L(x_i^{t+1}) < L(p_i^{\text{best}})$ **then** update $p_i^{\text{best}} \leftarrow x_i^{t+1}$.
    if $L(p_i^{\text{best}}) < L(g^{\text{best}})$ **then** update $g^{\text{best}} \leftarrow x_i^{t+1}$.
  end for
  $t \leftarrow t + 1, SBF \leftarrow L(\mathbf{g}^{\text{best}})$.
end while
return Adversarial example $\mathbf{g}^{\text{best}}$.

hierarchical structure with a maximum depth of 6 levels. We utilize the validation subset of the dataset[1], which consists of 17,927 audio files, each potentially containing multiple audio events. To evaluate model robustness, we treat the dataset as a one-class label classification task by grouping the audio files according to the top-level categories of the dataset's ontology. Table I presents these categories and their corresponding class distributions. We observe that some models can be easily fooled to classify an audio event as there is no event present. For this reason, we exclude "Silence" from the "Source-Ambiguous Sounds" category and treat it as a separate hypercategory.

TABLE I
CLASS DISTRIBUTION IN TOP-LEVEL CATEGORIES OF THE AUDIOSET VALIDATION SUBSET.

Category	Number of Samples
Natural sounds	306
Silence	6
Sounds of things	3,389
Channel, environment and background	240
Animal	990
Source-ambiguous sounds	788
Music	3,266
Human sounds	1,582

Environmental Sound Classification (ESC-50) [24] is a balanced dataset consisting of 2000 environmental recordings, each 5 seconds in duration, spanning 50 distinct audio event classes. These 50 classes are evenly distributed across 5 hypercategories: Animals, Exterior/Urban Noises, Natural Soundscapes & Water Sounds, Interior/Domestic Sounds, and Human Non-Speech Sounds.

B. Threat Model

We perform untargeted attacks with the goal of causing the model to misclassify the hypercategory. We filter out all samples that are misclassified by the model and retain only the correctly predicted ones. Our experiments operate in a black-box setting, where the adversary can only query the model and access the posterior probabilities. We vary the SNR value S in the set $\{5, 10, 15, 20, 25, 30\}$ dB, aiming to minimize the loss function L. We evaluate the robustness of three SOTA transformer-based models trained on Audioset.

Audio Spectrogram Transformer (AST) [16] is a transformer-based model designed for audio classification tasks with 88,132,063 trainable parameters. It process audio signals by first converting them into spectrograms, which are then divided into overlapping patches. These patches are projected into 1-dimensional embeddings, forming the input sequence to a series of stacked transformer blocks.

The Patchout faSt Spectrogram Transformer (PaSST) [17] extends AST and applies patchout to drop a portion of the input sequence. This method reduces training time and serves as a data augmentation, further improving the performance. This model consists of 86,153,759 parameters.

Bidirectional Encoder representation from Audio Transformers (BEATs) [18] is an iterative audio pre-training framework where an acoustic tokenizer and an audio self-supervised model are optimized by iterations. This is the first work in the audio domain that introduces an audio pretraining framework with discrete label prediction loss instead of reconstruction loss. BEATs achieves the highest performance of the three models containing 90,717,055 trainable parameters.

C. Implementation Details

Our implementation is written in Python, utilizing PyTorch as the deep learning framework. We source the pretrained models on AudioSet from their official repositories[2],[3],[4]. We evaluate the models on the hypercategory-level classification task by mapping each class label to its corresponding hypercategory. The adversarial attacks are performed on the subset of samples that are correctly classified by the model. For the ESC-50 dataset, we employ a 5-fold cross-validation approach. Specifically, we train a multilayer perceptron (MLP) classifier, with two hidden layers of size 512 and 256, using each of the three models as feature extractors excluding their classifier head. We evaluate each model on the held-out fold and performed adversarial attacks on the correctly classified samples within that fold. This process is repeated for all folds. For PSO we use a swarm of size $P = 25$ and a total of iterations $N = 20$. The inertia weight w is set to 0.9. The influence parameters

[1]https://www.kaggle.com/datasets/zfturbo/audioset-valid

[2]https://github.com/YuanGongND/ast

[3]https://github.com/kkoutini/PaSST

[4]https://github.com/microsoft/unilm/tree/master/beats

c_1, c_2 are equal to 1.2, and the initialization factor W is set to 0.5.

IV. Results and Discussion

A. Performance

The results in Table II demonstrate the performance of BEATs, PaSST, and AST on the AudioSet and ESC-50 datasets for the hypercategory classification task. On ESC-50, the mean average accuracy and mean F1 across the 5 folds are reported. As evident, BEATs consistently outperforms the other models, achieving the highest accuracy and F1 scores on both datasets. These results align with existing literature, where BEATs is reported as the best-performing model among the three for sound event detection tasks. Furthermore, on ESC-50, BEATs achieves 97% accuracy without hyperparameter tuning, which is consistent with the result of 98.1% reported in the original paper [18]. Finally, AST ranks second, surpassing PaSST on both datasets.

TABLE II
Model Performance on AudioSet and ESC-50.

Model	AudioSet		ESC-50	
	Accuracy	F1	Accuracy	F1
BEATs	**0.78**	**0.56**	**0.97**	**0.98**
PaSST	0.70	0.51	0.94	0.94
AST	0.77	0.53	0.96	0.96

B. Adversarial Robustness

Fig. 2 illustrates the robustness of the three models against adversarial attacks at varying SNR levels. The success rate, defined as the ratio of successful attacks, serves as the evaluation metric; higher success rates indicate lower robustness. BEATs is the least robust model across both datasets, with success rates significantly higher than those of PaSST and AST. For low SNRs ($\leq$ 15 dB) the success rate remains above 80% for AudioSet, and 40% ESC-50. At 20 SNR dB, where the perturbation is barely noticeable, the success rate is 68.03% for AudioSet, and 31.72% for ESC-50. A notable observation is the robustness gap for BEATs across the two datasets, where the success rates on AudioSet are significantly higher than those on ESC-50. By inspecting the distribution of adversarial examples across the hypercategories, we observe that most adversarial examples are classified as "Silence" by BEATs. This behavior may be attributed to the self-supervised training approach used for BEATs, which differs from the supervised training of PaSST and AST. During self-supervised training, BEATs learns acoustic tokenizers through a distillation process from a teacher model. If an audio event is not present during this process, the acoustic tokenizers may not be linked to that event, making it harder for the classifiers in the downstream procedure to learn a convex decision boundary for the event. In contrast, both PaSST and AST exhibit similar robustness across the two datasets. A minor difference is observed in the ESC-50 dataset, where PaSST appears slightly more robust than AST. The only distinction between these two models is that PaSST employs patchout during training, which acts as a regularization mechanism. This difference may help prevent overfitting to a specific data distribution, potentially leading to improved adversarial robustness.

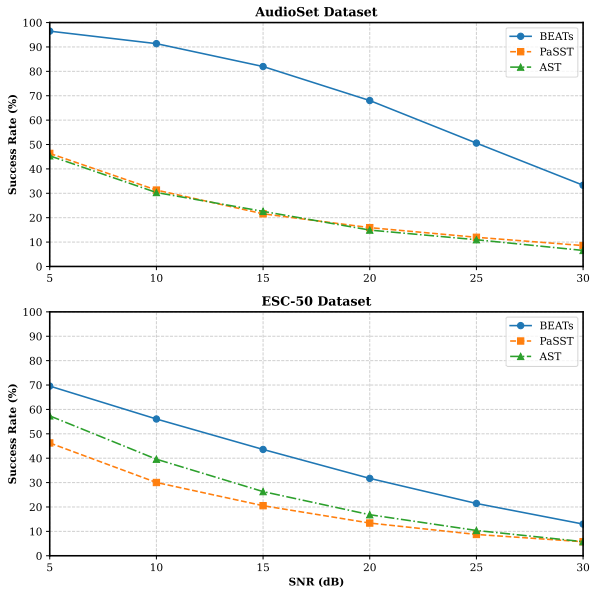

Fig. 2. The attacking success rate for the three models on AudioSet and ESC-50 for SNRs in $\{5, 10, 15, 20, 25, 30\}$ dB. Success rate is inversely propotional to robustness.

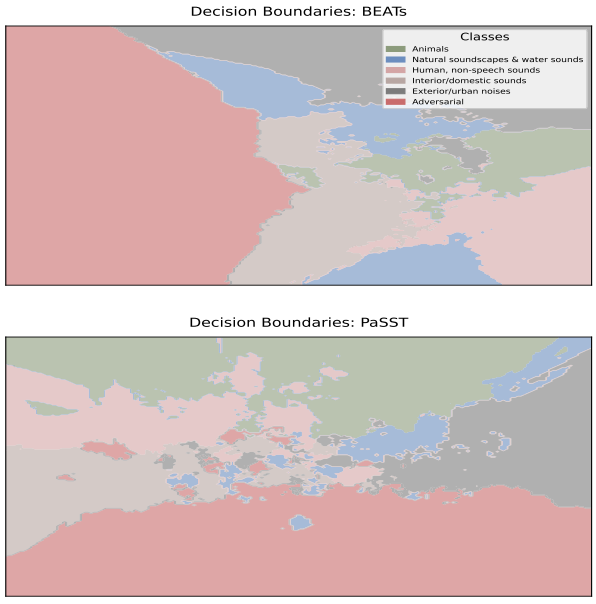

Fig. 3. The decision boundaries of the embeddings projected in 2D with PCA.

C. Further Discussion

As demonstrated in the previous results, PaSST exhibits the highest robustness against adversarial attacks. This finding contrasts with the results in Table II, where PaSST achieves the

lowest classification performance on clean (non-adversarial) data. This discrepancy suggests that robustness against adversarial attacks may be uncorrelated with model performance. We attribute this observation to the fact that as a model's performance on clean data increases, so does its tendency to overfit to a specific data distribution. Consequently, even small perturbations to the input may lead to incorrect predictions. Fig. 3 illustrates the BEATs and PaSST decision boundaries on ESC-50, with the embeddings projected in 2D via PCA. We visualize adversarial regions by attacking all samples and projecting the embeddings of the successful ones. We observe that adversarial examples occupy regions in the embedding space that are distinct from those explored during the training phase. These adversarial regions are clearly separated from the regions corresponding to the original classes, indicating that few, if any, embeddings of the original classes are mapped to these areas. This separation can be explained by the fact that deep neural networks, while theoretically continuous functions, behave like discrete functions in practice due to the inherently discrete nature of their training process. As a result, small perturbations in the input space can lead to representations that are significantly distant from each other in the embedding space.

V. Conclusions and Future Work

In this paper, we assessed the robustness of three SOTA transformers for sound classification against black-box adversarial attacks. We proposed a method to generate attacks with a fixed SNR, allowing robustness evaluation across varying imperceptibility levels. Our experiments show that while deep neural networks achieve exceptional performance, they remain highly vulnerable to adversarial examples, posing significant security risks for real-world deployment. It is still unclear why these models are susceptible to such inputs and whether there is a way to suppress these vulnerabilities. For future research, we aim to focus on answering these questions and, ultimately, develop methods to enhance the robustness of these models.

VI. Aknowledgements

This work is an outcome of the FaRADAI Project (https://faradai.eu). The authors gratefully acknowledge the support and funding provided by the European Union.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union nor the European Commission. Neither the European Union nor the granting authority can be held responsible for them.

References

[1] C Szegedy. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[2] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[3] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pages 372–387. IEEE, 2016.

[4] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017.

[5] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519, 2017.

[6] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017.

[7] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *International conference on machine learning*, pages 2137–2146. PMLR, 2018.

[8] Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE security and privacy workshops (SPW)*. IEEE, 2018.

[9] Yao Qin, Nicholas Carlini, Garrison Cottrell, Ian Goodfellow, and Colin Raffel. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In *International conference on machine learning*, pages 5231–5240. PMLR, 2019.

[10] Xingyu Zhang, Xiongwei Zhang, Meng Sun, Xia Zou, Kejiang Chen, and Nenghai Yu. Imperceptible black-box waveform-level adversarial attack towards automatic speaker recognition. *Complex & Intelligent Systems*, 9(1):65–79, 2023.

[11] Sajjad Abdoli, Luiz G Hafemann, Jerome Rony, Ismail Ben Ayed, Patrick Cardinal, and Alessandro L Koerich. Universal adversarial audio perturbations. *arXiv preprint arXiv:1908.03173*, 2019.

[12] Vinod Subramanian, Emmanouil Benetos, Ning Xu, SKoT McDonald, and Mark Sandler. Adversarial attacks in sound event classification. *arXiv preprint arXiv:1907.02477*, 2019.

[13] Tianyu Du, Shouling Ji, Jinfeng Li, Qinchen Gu, Ting Wang, and Raheem Beyah. Sirenattack: Generating adversarial audio for end-to-end acoustic systems. In *Proceedings of the 15th ACM Asia conference on computer and communications security*, pages 357–369, 2020.

[14] Vinod Subramanian, Arjun Pankajakshan, Emmanouil Benetos, Ning Xu, SKoT McDonald, and Mark Sandler. A study on the transferability of adversarial attacks in sound event classification. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 301–305. IEEE, 2020.

[15] Achyut Mani Tripathi and Aakansha Mishra. Adv-esc: Adversarial attack datasets for an environmental sound classification. *Applied Acoustics*, 185:108437, 2022.

[16] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021.

[17] Khaled Koutini, Jan Schlüter, Hamid Eghbal-Zadeh, and Gerhard Widmer. Efficient training of audio transformers with patchout. *arXiv preprint arXiv:2110.05069*, 2021.

[18] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. Beats: Audio pre-training with acoustic tokenizers. *arXiv preprint arXiv:2212.09058*, 2022.

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

[20] James Kennedy and Russell Eberhart. Particle swarm optimization. In *Proceedings of ICNN'95-international conference on neural networks*, volume 4, pages 1942–1948. ieee, 1995.

[21] Rainer Storn and Kenneth Price. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 11:341–359, 1997.

[22] Jon Vadillo and Roberto Santana. On the human evaluation of audio adversarial examples. *arXiv preprint arXiv:2001.08444*, 2020.

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

[24] Karol J Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018, 2015.