ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3315-6701-9/26/$31.00 ©2026 IEEE | DOI: 10.1109/ICASSP55912.2026.11460479

AUDIO CLASSIFICATION MODELS ARE VULNERABLE TO FILTER PERTURBATIONS

Justin Dettmer[⋆] *Annelot Bosman*[†] *Igor Vatolkin*[⋆] *Holger Hoos*[⋆†]

[⋆] Chair for Artificial Intelligence Methodology, RWTH Aachen University, Germany
[†]Leiden Institute of Advanced Computer Science, Leiden University, The Netherlands

ABSTRACT

Deep learning models are known to suffer from a lack of robustness to adversarial attacks. While the majority of research on the robustness of neural networks stems from the image classification domain, audio classifiers have also come under scrutiny, as their use in tasks such as speech processing is increasing. We introduce a novel adversarial attack method that creates band-pass filters as perturbations for mel spectrograms. These perturbations are closer to real-life conditions than conventionally created adversarial examples, as they can simulate the use of different recording devices or other acoustic properties. We demonstrate that state-of-the-art audio classifiers lack robustness against these adversarially-created filters, and we show that adversarial training with our attack method can help mitigate this effect on the widely studied NSynth, ESC-50 and SpeechCommands datasets.

***Index Terms*—** Audio Classification, Neural Networks, Adversarial Attacks, Audio Filters

1. INTRODUCTION

Neural networks form the basis for most state-of-the-art audio classification approaches. With millions to billions of trainable parameters, these models are very powerful; at the same time, they are vulnerable to often imperceptible changes, commonly referred to as *perturbations*, to originally correctly classified inputs, which can pose serious challenges when they are used in real-world applications. An ideal model should produce correct class predictions when presented with audio that may stem from different recording devices, contain background noise, or come from various acoustic environments.

A common way for assessing the robustness of machine learning models is to expose them to adversarial attacks [1]. These attacks work by slightly perturbing a given input, such as an audio recording, in a way that the model misclassifies it. In the audio classification domain, these adversarial examples have traditionally been crafted through noise that is added to the waveform [2]. Though these perturbations are successful in misleading models, they are generally not likely to occur naturally and are judged as artificial by human subjects [3].

In most non-security critical application contexts, there is no expectation that an attacker would attempt to mislead a model through adversarial methods. However, adversarial attacks may still be used to measure the robustness of a model, especially in the case of perturbations that are more likely to occur naturally. In this work, we introduce a novel method for creating perturbations as band-pass filters that are capable of approximately modelling acoustic properties such as the use of different microphones or recording setups.

We show that our modification of the well-established Projected Gradient Descent (PGD) attack [4] outperforms random search at finding filter-based adversarial examples. We further demonstrate that state-of-the-art audio classifiers are highly susceptible to these filter-based perturbations. Finally, we apply adversarial training in order to improve the robustness of our models on three common audio classification tasks.

In the following, we first provide an overview of related work, before explaining our novel modifications to PGD. Then, we will introduce our experiment setup, present results and summarise our key takeaways and propose suggestions for future work.

2. RELATED WORK

Researchers have long been concerned with the study of adversarial attacks, as reflected by the volume of research on methods for finding adversarial examples [1], training methods to harden neural networks against small perturbations [4] and formal verification of the robustness of neural networks [5]. Most of this research is focused on the classification of images.

While the community has mostly considered local robustness properties—as this problem is already hard to solve—various works applied adversarial attacks on semantically meaningful visual perturbations, such as Fourier-based visual filters [6].

Hanspal *et al.* [7] defined properties for image classifiers based on Fourier transforms, which allowed them to construct properties with interdependency between pixels and perform verification for those properties using existing neural network verifiers. Our work similarly considers a robustness property

based on filters, where "pixels" of the spectrogram are not independently perturbed.

In the audio domain, some work considering robustness against adversarial attacks has been published over the last few years. Conventional methods for reducing the impact of adversarial examples on the accuracy of neural networks include local-smoothing and downsampling [8].

Carlini *et al.* [2] were the first to show that targeted adversarial attacks are possible in the audio domain by adding humanly imperceptible, carefully designed changes to a waveform. These small changes make it possible for a speech recognition classifier to be fooled into predicting any desired outcome.

Du *et al.* [9] proposed an adversarial attack method called SirenAttack that is designed for audio tasks such as speech recognition. They combined the commmonly used gradient approach with a heuristic algorithm for generating perturbed samples.

Mauch *et al.* [10] used various audio degradation techniques, including high- and low-pass filters, to evaluate the robustness of audio processing algorithms for a selection of music information retrieval tasks. While the inclusion of filters in their work creates a close connection to ours, they did not apply their toolbox to machine learning methods.

Closely related to ours is the work by Sturm *et al.* [11], which studied different waveform augmentations in a musical genre classification task. These augmentations were more semantically meaningful, echoing our motivation to create more realistic evaluation conditions for audio models, however did not include the aspect of adversarially crafting perturbations to measure the robustness of a model.

3. FILTER PERTURBATIONS

In this section, we will provide background information on PGD and explain the modifications we made to it in order to generate filter-based perturbations.

3.1. Projected Gradient Descent

Madry *et al.* [4] proposed using PGD to train neural networks to be more robust against adversarial perturbations. PGD is a multi-step extension of the Fast Gradient Sign Method (FGSM) [1], incorporating random initialisation within the ε-ball. The ε-ball, in the case of image classification, is defined as the range in which a perturbation may lie, as given by the l_∞ distance between the original and the perturbed image.

FGSM and PGD both generate adversarial examples by propagating an instance through a pretrained neural network and applying a min–max optimisation strategy to modify the originally correctly classified sample in order to maximise the loss, thereby increasing the probability of misclassification. During adversarial training, these generated adversarial examples are included in the training process, forcing the network to minimise the loss on both clean and perturbed samples, which improves robustness.

3.2. Filter-based PGD

Formally, a neural network audio classifier can be described by a function f_θ in $\mathbb{R}^{N \cdot T} \rightarrow \mathbb{R}^m$, where θ is the set of trained parameters for f_θ, x is an input spectrogram with N frequency bands and T time frames, and m is the number of possible output classes.

In this work, we define a filter as a vector h, with N elements as given by the size of the input spectrogram. We assume that if a filter influences a frequency band $n \leq N$, it influences the frequency with a constant magnitude over time, creating a discrete band-pass filter. This means that a frequency is either amplified or attenuated at the same rate over time, similar to how the resonant frequency of a room is not expected to change within a short timeframe. In the case of $\forall n : \ h_n = 1$, the filter represents the identity function, making no change to the spectrogram at all. We initialise our algorithm with this particular filter, which we denote by h^0.

Considering our initial filter h^0, we define $S_{p,\varepsilon}$ as the region around h^0 encapsulating all feasible filters given by

$$\lim_{p \to \infty} S_{p,\varepsilon} = \{h : ||h - h^0||_p \leq \varepsilon\}, \tag{1}$$

remaining consistent with existing robustness literature by using the l_∞ norm. If we find a filter in the region $S_{p,\varepsilon}$ that leads to misclassification, we consider this as a *filter-based adversarial example.* Note that multiplication of the filter with the spectrogram in the frequency domain is equivalent to a convolution in the time domain [12, pp. 39–41].

We adapt the aforementioned PGD algorithm to generate filter-based perturbations according to our definition. At any step t, a filter h^t for an input spectrogram x is iteratively computed in the following way, until a filter-based adversarial example is found or until the predefined number of steps is reached:

$$h_i^t = \mathit{clip}\left(h_i^{t-1} + \alpha \cdot \mathit{sign}\left(\sum_{j=1}^{T} \left(\nabla_{x^{t-1}} L\left(\theta, x^{t-1}, y\right)\right)_{i,j}\right)\right) \tag{2}$$

where $x^t = h^t \cdot x$ is the candidate perturbed spectrogram at step t, L is a cross-entropy loss function and α is a constant step size. Here, *clip* bounds each element of the resulting filter h^t between $[1 - \varepsilon, 1 + \varepsilon]$, and the summation over the time dimension of the spectrogram aggregates the loss terms.

4. EXPERIMENTS

We demonstrate the effect of filter-based adversarial attacks against a state-of-the-art Patchout Spectrogram Transformer (PaSST) [13] and the commonly-used CNN14 model [14] on three audio classification datasets.

ESC-50 [15] is a dataset for environmental sound classification for 50 classes of sounds from categories such as animal noises, city soundscapes and domestic sounds.

SpeechCommands [16] contains recordings of 35 words for the task of keyword detection, recorded with phone or laptop microphones.

NSynth [17] consists of 305 979 musical instrument samples of varying pitches, sourced from commercially available sample libraries. We use this dataset for the task of musical instrument recognition.

We used the weights of PaSST provided by Koutini *et al.* [13] and by Kong *et al.* [14] for CNN14, which were both pretrained on AudioSet [18]. For every dataset, we trained five models each on different random seeds with the Adam optimiser, learning rates of 10^{-5} for PaSST and 10^{-4} for CNN14, and early stopping based on the accuracy on the validation set with a patience period of ten epochs. Due to the size of the dataset, the validation accuracy on NSynth was evaluated after 20% of training batches were used in every epoch, and the learning rate was halved as the validation accuracy did otherwise converge too quickly for early stopping to choose an effective stopping epoch.

All datasets were augmented with SpecAugment [19], random rolling and gain attenuation/amplification, and two-level mixup [20].

The trained models were then evaluated by performing filter-based PGD attacks with multiple ε values, and we measured the classification accuracy on the perturbed test sets. To assess the efficacy of our PGD adaptation, we further compared it to a random search within the same ε-ball.

In a subsequent experiment, we applied adversarial training to PaSST on each of these datasets with the same set of hyperparameters. One model was trained for every ε in the range of $(0, 1]$ in equidistant steps of size 0.1. Finally, all adversarially trained models were evaluated on their respective test sets with the same filter-based PGD attack described above.

5. RESULTS AND DISCUSSION

In this section, we will present and analyse the results of our experiments and discuss what they show about the robustness of audio models regarding filter-based attacks.

5.1. Attacks on Baseline Models

The adversarial accuracy of our baseline models is shown in Fig. 1. For all three datasets, it is clear that the trained baseline models are highly susceptible to filter-based adversarial attacks, even for small ε values. We see that for ESC-50 and SpeechCommands, our filter-based PGD performs significantly better than random search at finding adversarial examples. Using a one-sided Wilcoxon signed-rank test, we confirmed statistical significance ($p < 0.05$) for all model and

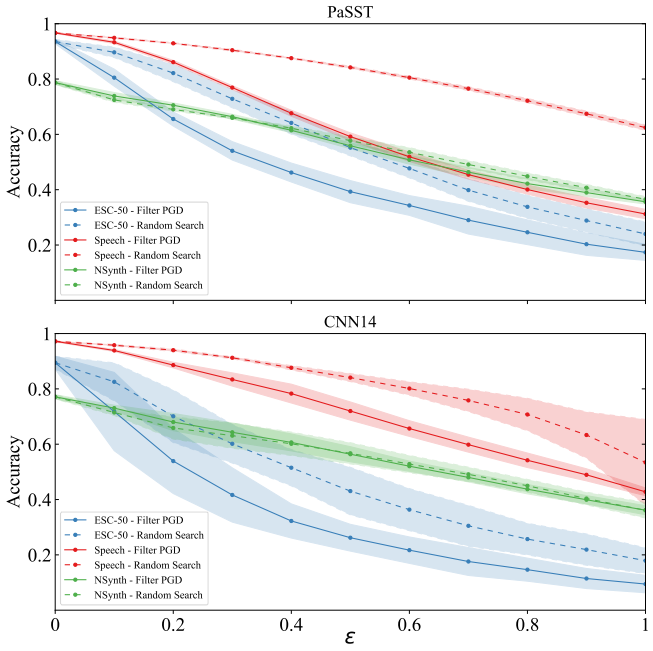

Fig. 1. Mean test accuracy aggregated for five trained models against filter-based PGD and random search. 95% confidence intervals are shown around the plotted means for all three datasets.

dataset combinations except for CNN14 on NSynth, where the null hypothesis could not be rejected. This shows that the loss information of individual spectrogram values successfully propagates across the summation of signs in Eq. 2 and that it shifts the search into more promising regions. However, given that for $\varepsilon = 1$ there is a theoretical global optimum, where every element of the filter is set to 0, we can see that neither method is able to find this particular optimum reliably. It may be the case that finding this filter is akin to a needle-in-a-haystack problem, where no global trend could guide the search process towards the global optimum.

A further consideration in favour of our approach is that the number of mutable parameters is vastly smaller than that of conventional adversarial attacks. While a conventional attack may be able to modify every element of the spectrogram or waveform individually, our filter approach limits this to the number of mel bands in the spectrogram, in our case 128 for PaSST and 64 for CNN14. Despite this limitation, the significant degradation of classification accuracy across all datasets shows that, despite various data augmentation methods, state-of-the-art deep leaning models are not even robust to this theoretically weaker adversary.

We provide listening examples alongside our code on GitHub[1] to show that, especially for small ε, the filter-based perturbations are not easily noticeable. As our constant band-

[1]https://github.com/ADA-research/AdvFilters

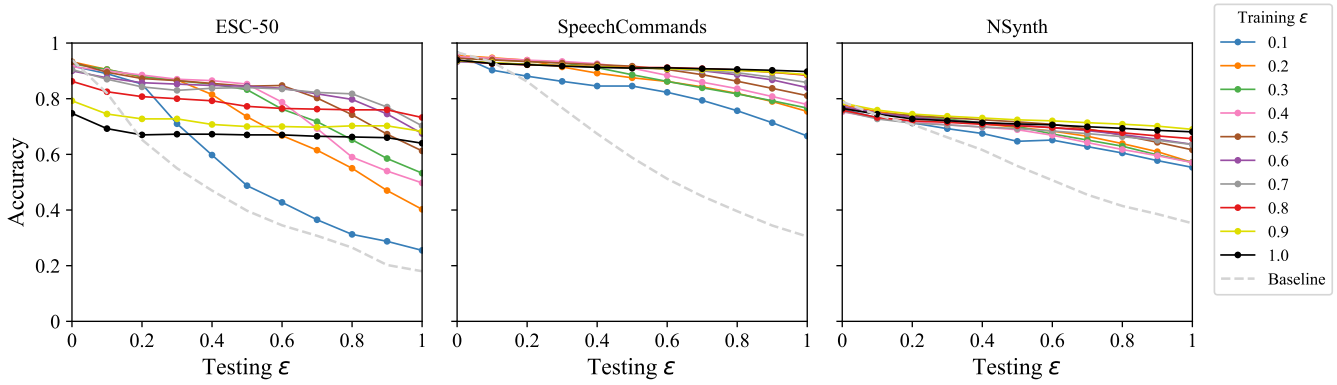

Fig. 2. Test accuracies of PaSST models against filter-based PGD after adversarial training with different ε values. The x-axis shows the ε with which the attacks on the test set were performed, while the line colour indicates the ε with which it was trained.

pass filters only affect the frequency content of the spectrogram, but leave temporal structures such as rhythmic noises intact, the perturbed audio sounds more natural to the human ear than conventionally created adversarial examples.

5.2. Adversarial Training

In our second experiment, we applied adversarial training in an attempt to improve the robustness of these models. As shown in Fig. 2, adversarial training creates models that show much better accuracy against filter-based PGD, even when trained with very small ε values. While the robustness of each model still declines once the attack is performed with an ε higher than the one that a model was trained with, we did observe improved robustness over the baseline models across all datasets. There is a slight trade-off between the adversarial accuracy at higher ε values and the accuracy on the unperturbed test set. This trade-off is known [4], yet remains within a reasonable range for the models trained with smaller ε. One should consider the application contexts in which a model may be deployed, and choose a suitable ε value.

Attacks on the ESC-50 dataset remain the most successful among the three datasets we tested. This may be correlated with the comparatively small size of the dataset, containing only 1200 training samples after removing the testing and validation folds. The wider spread in the 95% confidence intervals in Fig. 1 further shows that the small dataset results in more inconsistent model performance compared to the larger two datasets.

It is important to note that the robustness gains of adversarial training come at the cost of a substantial increase in training time. Especially on the larger datasets NSynth and SpeechCommands, we noticed an up to 10-fold[2] increase, due to the sheer number of adversarial attacks that had to be performed on every batch.

[2]Experiments were run on a compute cluster with NVIDIA H100 GPUs.

6. CONCLUSION

In this work, we have introduced a novel approach to creating adversarial examples for neural network audio classifiers. With our filter-based method, we were able to show that state-of-the-art deep learning classifiers are highly vulnerable to slightly filtered mel spectrograms on three commonly used audio classification datasets from different domains. We further showed that, through adversarial training, this problem can be partially mitigated, albeit at the cost of a minor trade-off in accuracy on the unperturbed test set.

As, to our knowledge, this is the first work investigating such filter-based attacks, there are many avenues for future work. To get an even better idea of the effect of real-world filters, it would be interesting to restrict the range of available adversarial filters to a discrete set that is specifically crafted to model particular recording devices or acoustic environments. Further, our approach could be modified to allow for a smaller number of bands, as is the case in more common band-pass filters, such as the ones used in FilterAugment [21]. In that context, it may also make sense to bound our attack by using the signal-to-noise ratio, as is commonly done in noise-based attacks.

The use of neural network verification tools to give robustness guarantees with regards to filters would be another important contribution, as adversarial robustness alone gives only an upper bound on the accuracy of a model under adversarial conditions.

Lastly, we believe a human study in the style of Vadillo *et al.* [3] would provide valuable insight into the perceptibility of our perturbations compared to noise-based attacks.

ACKNOWLEDGEMENTS

Holger H. Hoos gratefully acknowledges support through an Alexander-von-Humboldt Professorship in Artificial Intelligence. Computational resources were provided by the German AI Service Center WestAI.

REFERENCES

[1] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial examples," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.

[2] Nicholas Carlini and David A. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *2018 IEEE Security and Privacy Workshops, SP Workshops*. 2018, pp. 1–7, IEEE Computer Society.

[3] Jon Vadillo and Roberto Santana, "On the human evaluation of universal audio adversarial perturbations," *Computers & Security*, vol. 112, pp. 102495, 2022.

[4] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2018.

[5] Linyi Li, Tao Xie, and Bo Li, "SoK: Certified robustness for deep neural networks," in *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2023, pp. 1289–1310.

[6] Xiu-Chuan Li, Xu-Yao Zhang, Fei Yin, and Cheng-Lin Liu, "F-mixup: Attack cnns from fourier perspective," in *Proceedings of the 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 541–548.

[7] Harleen Hanspal, Alessandro De Palma, and Alessio Lomuscio, "Robustness to perturbations in the frequency domain: Neural network verification and certified training," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025, pp. 591–600.

[8] Guangke Chen, Zhe Zhao, Fu Song, Sen Chen, Lingling Fan, Feng Wang, and Jiashui 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.

[9] 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*, 2020, pp. 357–369.

[10] Matthias Mauch and Sebastian Ewert, "The audio degradation toolbox and its application to robustness evaluation," in *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR)*, 2013, pp. 83–88.

[11] Bob L Sturm, "The "horse" inside: Seeking causes behind the behaviors of music content analysis systems," *Computers in Entertainment (CIE)*, vol. 14, no. 2, pp. 1–32, 2017.

[12] Meinard Müller, *Information Retrieval for Music and Motion*, Springer-Verlag Berlin Heidelberg, 2007.

[13] Khaled Koutini, Jan Schlüter, Hamid Eghbal-zadeh, and Gerhard Widmer, "Efficient training of audio transformers with patchout," in *Proceedings of Interspeech 2022*, pp. 2753–2757.

[14] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[15] Karol J. Piczak, "ESC: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM Conference on Multimedia (ACM MM)*, 2015, pp. 1015–1018.

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

[17] Jesse H. Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan, "Neural audio synthesis of musical notes with wavenet autoencoders," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*. 2017, vol. 70 of *Proceedings of Machine Learning Research*, pp. 1068–1077, PMLR.

[18] Jort F. Gemmeke, Daniel P.W. 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 *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[19] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Proceedings of Interspeech 2019*, pp. 2613–2617.

[20] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2018.

[21] Hyeonuk Nam, Seong-Hu Kim, and Yong-Hwa Park, "Filteraugment: An acoustic environmental data augmentation method," in *Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 4308–4312.