***Defense for Black-box Attacks on Anti-spoofing Models by Self-Supervised Learning***

High-performance anti-spoofing models act as protective shields for ASV systems (Automatic Speaker Verification) by using advanced ML and signal processing techniques to differentiate between genuine human speech and spoofed/fake speech signals. They are trained on large datasets to learn patterns and features that distinguish between the two. Adversarial attacks are malicious attempts to deceive machine learning pipelines, and the Mockingjay model employs self-supervised learning to protect anti-spoofing models against such attacks in black-box scenarios. The Layerwise Noise-to-Signal Ratio (LNSR) is a metric used to evaluate the effectiveness of deep learning models in countering adversarial noise. The LNSR analysis reveals that the self-supervised learning model attenuates adversarial noise layer by layer, indicating its ability to mitigate the effects of such attacks.

Concept of Adversarial attacks was proposed by Szegedy et al showing that deep neural networks used in computer vision tasks are vulnerable to attacks. Small changes applied on input samples results in incorrect model predictions.

Proactive Defense: Train new models to counter the adversarial attacks. Adversarial Training incorporates adversarial examples into training data to enhance robustness. It is time consuming and may not cover all possible attack algorithms.

Passive Defense: defend against the attacks without modifying the model. Gaussian, Median, or Mean filters help mitigate the effects of deliberately crafted adversarial attacks.  
Self-Supervised models can act as deep filters, extracting information from spoofed input to counter adversarial attacks.

***ADVERSARIAL ATTACKS: -***

Adversarial example x’=x+ δ Adversarial noise

Original example

Different strategies for searching δ result in different attack algorithms. The fast gradient-sign method (FGSM) and the projected gradient descent method (PGD) are used in this paper.

Black-box scenario: attacker and target model are different, attacker doesn’t have access to inner parameters of target model, but the attacker gathers input-output pairs from the target model to train a substitute model. The substitute model is then used to generate adversarial examples that can deceive the target model, despite the attacker lacking full knowledge of its internals.

White-box scenario: attacker has complete knowledge of the target model allowing for more direct and informed attacks.

***MOCKINGJAY: -***

It uses a multi-layer transformer encoder with self-attention capture dependencies between different time frames in speech data and a feed forward prediction to reconstruct the masked frames. This combination enables the model to learn robust speech representations for tasks like anti-spoofing.

***SELF-SUPERVISED LEARNED ADVERSARIAL DEFENDER: -***

Here, Mockingjay is proposed to protect anti-spoofing models. Traditional features used in speech analysis may miss important information. Mockingjay extracts richer information from speech and makes it available to the anti-spoofing model.

In a black-box attack, where attackers do not know about Mockingjay, they try to add adversarial noise to the input spectrograms. However, Mockingjay reduces the effect of this noise and prevents its transferability, thereby countering the attack.

The effectiveness of Mockingjay comes from its ability to weaken and extract key information from noisy spectrograms, including adversarial noise. Additionally, the different training approaches used in Mockingjay and the attacking model contribute to its defense capabilities.

***LAYERWISH NOISE TO SIGNAL RATIO: -***

LNSR helps us understand how much of the attack signal is present in each layer. If LNSR decreases as we move to higher layers, it indicates that Mockingjay is effectively reducing the impact of the attacking noises.

***EXPERIMENT: -***

The study uses a dataset of fake audios and employs the Mockingjay model, LCNN, and SENet for anti-spoofing. Adversarial examples are generated using basic LCNN and SENet as attackers and evaluated against the target models using FGSM and PGD attacks with varying epsilon values. The results emphasize that the proposed Mockingjay defense mechanism effectively defends against adversarial attacks compared to other methods, including hand-designed filters and models trained from scratch. The pre-trained Mockingjay model significantly reduces LNSR, showcasing its effectiveness in mitigating adversarial noise, while the random parameterized Mockingjay has limited impact. This further emphasizes the crucial role of pre-training in the defense approach.

***REFERENCES: -***  
1. <https://www.analyticsvidhya.com/blog/2022/09/machine-learning-adversarial-attacks-and-defense/#:~:text=At%20its%20core%2C%20adversarial%20attacks,%2C%E2%80%9D%20which%20can%20elicit%20misclassification>.

2. <https://www.chat.openai.com>

3. <https://www.isca-speech.org/archive_v0/Interspeech_2020/pdfs/2026.pdf>

4. <https://en.wikipedia.org>

|  |  |  |  |
| --- | --- | --- | --- |
| Detect/Mitigate | Universal | Prior Knowledge | Only Defensive? |
| Does not directly detect or modify changes | Against multiple attacks. | It isn’t needed, as Mockingjay uses the info present in speech data | It’s primarily designed and used as a defense mechanism. |

|  |  |  |  |
| --- | --- | --- | --- |
| Add Ons | Modified Network | Test on Attack Knowledge | Attack |
| It’s a standalone model | Rather than modifying the network/architecture, the main focus is to use it as a defense model. | Both scenarios were used (Blackbox and Whitebox method) | Fast Gradient Sign Method (FGSM) and PGD (Projected Gradient Descent) |

|  |  |  |
| --- | --- | --- |
| Architecture | Data Set | Accuracy |
| 12 layers of Transformer Encoders. | LA partition of the ASVspoof 2019 challenge dataset (fake audios generated by text-to-speech and voice conversion techniques) | Pre-trained Mockingjay successfully lowers the LNSR, but random parameterized model saturates at one point. |