applied sciences

MDPI

Article

IG-Based Method for Voiceprint Universal Adversarial Perturbation Generation

Meng Bi [1], Xianyun Yu [1], Zhida Jin [2] and Jian Xu [2,*]

[1] College of Software, Shenyang University of Technology, Shenyang 110178, China; bim@sut.edu.cn (M.B.)
[2] College of Software, Northeastern University, Shenyang 110819, China
* Correspondence: xuj@mail.neu.edu.cn

Abstract: In this paper, we propose an Iterative Greedy-Universal Adversarial Perturbations (IGUAP) approach based on an iterative greedy algorithm to create universal adversarial perturbations for acoustic prints. A thorough, objective account of the IG-UAP method is provided, outlining its framework and approach. The method leverages a greedy iteration approach to formulate an optimization problem for solving acoustic universal adversarial perturbations, with a new objective function designed to ensure that the attack has higher accuracy in terms of minimizing the perceptibility of adversarial perturbations and increasing the accuracy of successful attacks. The perturbation generation process is described in detail, and the resulting acoustic universal adversarial perturbation is evaluated in both target-attack and no-target-attack scenarios. Experimental analysis and testing were carried out using comparable techniques and dissimilar target models. The findings reveal that the acoustic generality adversarial perturbation produced by the IG-UAP method can obtain effective attack results even when the audio training data sample size is minimal, i.e., one for each category. Moreover, the human ear finds it difficult to detect the loss of original data information and the addition of adversarial perturbation (for the case of a target attack, the ASR values range from 82.4% to 90.2% for the small sample data set). The success rates for untargeted and targeted attacks average 85.8% and 84.9%, respectively.

Keywords: iterative greedy; universal adversarial perturbations; voiceprint; targeted attacks

Citation: Bi, M.; Yu, X.; Jin, Z.; Xu, J. IG-Based Method for Voiceprint Universal Adversarial Perturbation Generation. *Appl. Sci.* **2024**, *14*, 1322. https://doi.org/10.3390/app14031322

Academic Editor: Ugo Vaccaro

Received: 19 December 2023
Revised: 26 January 2024
Accepted: 30 January 2024
Published: 5 February 2024

1. Introduction

In recent years, the field of artificial intelligence has seen rapid development, and hardware equipment has substantially increased in arithmetic power. Consequently, current deep learning technology finds diverse applications in fields like computer vision [1], network security, medical analysis, computer graphics, and recommender systems. Specific scenarios such as speech recognition, image recognition, credit assessment, filtering malicious emails, resisting malicious code attacks, and cyber attacks have fostered the advancement of diverse fields and industries. Since the publication of the AlexNet convolutional neural network model by Alex Krizhevsky et al. in 2012 [2], deep learning models have made significant progress in classification effectiveness, surpassing traditional classification methods. Successive neural network models, including the ResNet model [3], the VGG model [4], and the GoogleNet model [5], have further improved classification performance, making deep learning models widely used in image classification and voice print recognition.

With the advancement of artificial intelligence, voiceprint recognition authentication has become an increasingly important aspect of our daily lives and work. Security concerns resulting from this are gaining significant attention. Signal processing and deep learning algorithm models have significantly enhanced the accuracy and reliability of voiceprint recognition in comparison to conventional speech recognition methods. However, although there are benefits, several security concerns have arisen.

Szeged et al. originally proposed the concept of adversarial samples in 2014 [6]. They discovered that neural network models can make classification errors due to adversarial perturbations. Furthermore, these neural network models, which have excellent classification performance, can also experience incorrect classification predictions when under attack from malicious adversarial samples. Moosavi-Dezfooli et al. introduced the concept of generic adversarial perturbation which can be used to deceive neural network models [7]. They developed a Greedy-based attack algorithm to target image classification models in untargeted scenarios. This technique presents various opportunities and possibilities for deception. This approach to producing usual adversarial perturbations involves normalizing them by combining all of the adversarial perturbations of each sample in the original dataset. Every targeted adversarial perturbation will shift the initial data to the decision boundary of the target classifier, lowering the target model's genuine category classification confidence. Furthermore, a significant association between the decision boundary's structure and generalized adversarial perturbations was presented by Moosavi-Dezfooli et al. [8]. In 2019, Paarth et al. were the first to demonstrate the presence of generalized adversarial perturbations in the audio domain [9], which do not relate to any specific audio sample in the dataset but can be added to any of the audio samples.

In this paper, an acoustic genericity adversarial perturbation generation method is designed based on the iterative greedy method. The acoustic features of audio samples are computed in the iterative process to generate an acoustic genericity adversarial perturbation, which improves the production efficiency of the adversarial perturbation and the efficiency of the adversarial attack.

2. Design Consideration

Iterative greedy is an efficient and widely-used approach for solving generic adversarial perturbations. This method progressively optimizes the objective function through iterative computation. At each iteration, the optimizer calculates and adjusts the objective function based on the current parameter settings to locate the optimal point of the objective function. This enables the optimization of the adversarial perturbation and attainment of the optimal adversarial perturbation solution. The acoustic features of audio samples are calculated during the iterative process of the IG-UAP generation method, enabling the creation of acoustic generality adversarial perturbation. This enhances both the production efficiency of the adversarial perturbation and the efficiency of the attack. Please see Figure 1 for the specific algorithm flow.

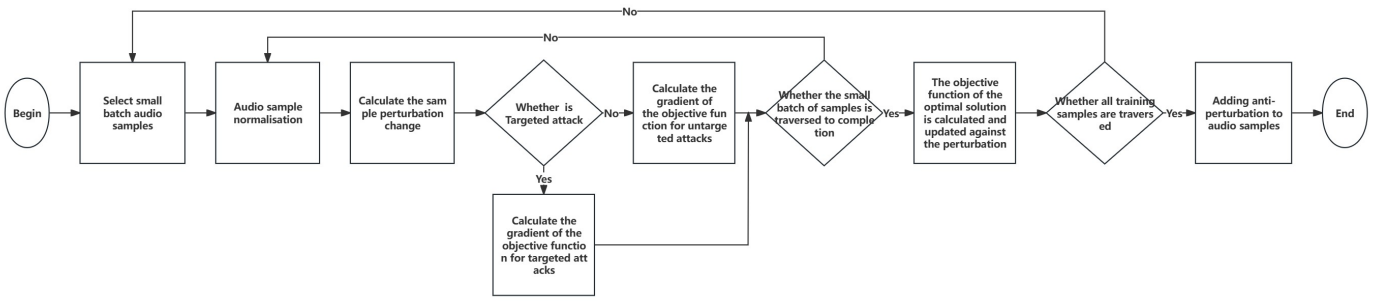


Figure 1. IG-UAP adversarial perturbations generation process.

In the context of noise perception within the acoustic pattern classification system, this paper proposes a novel objective function that employs authentic metrics, particularly the sound pressure level, to quantify the noise perception level as an optimization problem. At each iteration, a subset of audio data samples is chosen from the dataset and normalized. Then, the updated gradient is estimated by computing the degree of perturbation transformation for each training audio sample to effectively solve the objective function for that subset of samples. Through multiple rounds of iteration, the

objective function of the dataset is continually updated to yield the optimal solution. The attack success rate determines the stopping point for the iteration, and once the set threshold is reached, the final acoustic generality adversarial perturbation is produced. This perturbation is then added to the audio samples to create the adversarial samples.

3. Symbolic Description

Table 1 describes the symbols related to this paper.

Table 1. Notation of attack description.

Symbolic	Meaning
X	Original audio samples
t	The predicted label of the target that the attacker wants the target model to output
y_l	A category in untargeted attacks
$L(\cdot)$	Objective function
$G(\cdot)$	Updated gradient
δ	Expectation of deception rate
v'	The generalization of this dataset against perturbations
c	Penalty constant factor
K	Confidence level for controlling sample misclassification

4. IG-UAP Methodology

In the iterative greedy approach, the primary aim is to minimize the objective function of a batch of samples from the dataset to calculate the generalizability against perturbations. An iterative greedy-based algorithmic process is designed to generate an adversarial perturbation algorithm for acoustic generalization. The symbols used in the algorithm are described in detail, followed by an explanation and analysis of the specific role of each step. The following section elaborates on the iterative optimization process, including the solution steps and schemes of the optimization problem. Additionally, a new objective function is designed to implement an iterative greedy-based adversarial perturbation generation method for acoustic generality. Relevant experimental analyses are carried out under target-attack and targetless-attack scenarios.

This is because the objective function fundamentally characterizes the performance against such perturbations. In the context of perceiving noise in a sound pattern classifier, the degree of noise perception may be measured using a reliable metric, such as the sound pressure level (SPL), as demonstrated in Equation (1). Consequently, the SPL is implemented instead of the $\boldsymbol{L_p}$ paradigm for constraints. This metric is utilized as one of the objective functions in the optimization problem of this paper. The goal is to lessen the SPL of the perturbation, which is measured in decibels (dB). The problem of creating perturbations in a target attack can be reformulated as the following constrained optimization problem.

$$SPL(v) = 20logP(v) \tag{1}$$

$$P(v) = \sqrt{\frac{1}{N}\sum_{n=1}^{N} v_n^2} \tag{2}$$

The optimization problem presented in the equations can be solved with a gradient-based algorithm. As a result, a new parameter w must be introduced, which is defined in the equation. Audio example x_i must be transformed into the tan space, and then, the perturbation data can be converted into the effective range of [0, 1] using Equation (3). This ensures that the constraint of [0, 1] is met.

$$w_i = \frac{1}{2}(\tan(x_i' + v') + 1) \tag{3}$$

$$x_i' = \tan^{-1}((2x_i - 1) \times (1 - \varepsilon)) \tag{4}$$

$$v' = \tan^{-1}((2v - 1) \times (1 - \varepsilon)) \tag{5}$$

where ε is a small constant determined by the polarity of the transformed signal to prevent x_0' and v_0' from taking infinite values.

The formula mentioned above is simplified to an optimization problem to obtain the best possible solution. Additionally, in the event of a target attack, the formula is presented in (6).

$$w_{i_{min}} = \begin{cases} L(w_i, t) = SPL\left(\frac{1}{2}\ln\left(\frac{w_i}{1 - w_i}\right) - {x_i}'\right) + cG(w_i, t) \\ G(w_i, t) = max\{max\{f(w_i)_j\} - f(w_i)_t, -\kappa\} \end{cases} \tag{6}$$

where the target category is represented by t, the output of the presoftmax layer (logit) of the neural network for category j is represented by $f(w_i)_j$, a positive constant known as the "penalty coefficient" is represented by c, and the confidence level that controls the misclassification of samples is represented by κ. This formula enables the attacker to manage the confidence level of the attack. For untargeted attacks, we modify Equation (6) according to Equation (7).

$$w_{i_{min}} = \begin{cases} L(w_i, y_l) = SPL\left(\frac{1}{2}ln\left(\frac{w_i}{1 - w_i}\right) - {x_i}'\right) + cG(w_i, y_l) \\ G(w_i, y_l) = max\{f(w_i)_{y_l} - max\{f(w_i)_j\}, -\kappa\} \end{cases} \tag{7}$$

This is ultimately calculated using a gradient-based optimization algorithm, including the Adam algorithm, to minimize the losses defined in Equations (6) and (7). Various optimization algorithms, for example, AdaGrad, Standard Gradient Descent, Nesterov Momentum Gradient Descent and RMSProp, have been assessed, but Adam achieves convergence in fewer iterations and produces comparable outcomes.

5. Generic Adversarial Perturbation Generation Process

During the iterative optimization process, a subset of audio training samples will be selected for traversal in each round of iteration. Before proceeding, each input sample from the original audio data set will be normalized within the interval [0, 1] to adhere to constraints. Here, 0 indicates the minimum amplitude, and 1 represents the maximum amplitude. Subsequently, the perturbation signal transform of each audio sample will be computed to determine the objective function gradient based on the perturbation signal. After traversing all of the relevant samples, the optimization process will be complete. Combining the gradients obtained from solving the objective function optimally using Adam's rule, we calculate the generic adversarial perturbation. We then iteratively update the optimized perturbation to the original audio samples, ultimately generating the generic adversarial perturbation. We stop iterating and save the adversarial samples once the attack success rate reaches the preset threshold value. Details of the training process can be found in Algorithm 1.

Algorithm 1 IG-Based Method for Voiceprint Universal Adversarial Perturbation Generation

Input: Raw audio training Sets x_{train}, Training set labels y_{label}, Expectation of deception rate δ, Targeted attack category t

Output: Generalized counteracting perturbation v'

1: Initialize the generic adversarial perturbation v' to 0
2: Selecting small samples S in the dataset, $x_i \in (x_{train}, y_{label})$
3: Initialize the update gradient g to 0
4: Normalize the audio samples x_i to the interval range $[0,1]$, $x_i' = \tan^{-1}((2x_i - 1) \times (1 - \varepsilon))$
5: Calculate the acoustic perturbation signal transform for sample x_i:$w_i = \frac{1}{2}(\tan(x_i' + v') + 1)$
6: Calculate the gradient of the objective function. $g \leftarrow g + \frac{\partial L(\omega_i, t)}{\partial \omega_i}$ if it is a targeted attack and $g \leftarrow g + \frac{\partial L(\omega_i, y_i)}{\partial \omega_i}$ if it is an untargeted attack
7: Repeat steps 4 to 6 until each sample x_i in the small batch sample set S has been traversed
8: Based on the obtained update gradient g, which is updated according to Adam's rule, yields $\Delta v'$
9: Updating generalized counter perturbation $v' \leftarrow v' + \Delta v'$
10: Repeat steps 2 through 9 until $\mathrm{ASR}(X, V_G) > 1 - \delta$
11: Output generalized anti-perturbation v'

6. Experimental Analysis

Five pretrained voiceprint classification models, namely 1DCNN Rand, 1DCNN Gamma, ENVnet-V2, Sincnet, and SincNet+VGG19, are employed on the UrbanSound8k dataset [10], applying the IG-UAP method to evaluate their performance in untargeted and targeted attacks. Implementation and experimental comparisons were performed on four existing voiceprint universal adversarial perturbation generation methods, which include FGSM-UAP [11], PGD-UAP [12], C&W-UAP [13], and MSCW-UAP [14]. Table 2 displays the experimental environment. The sample sizes were small, so caution in drawing conclusions is advised.

Table 2. Test environment.

Components	Category	Item
Hardware	CPU	Xeon(R) W-2123 CPU 3.60 GHz
	Random Access Memory	16 GB 2133 MHz LPDDR3
	Video Card	NVIDIA GeForce RTX 2080 Ti12GB
Software	Operating System	Windows 10 64-bit System
	Programming Languages	Python
	Third Party Libraries	Pytorch, torchaudio, flask, bootstrap

FGSM-UAP combines the adversarial perturbation generated by FGSM with UAP, generating a universal perturbation applicable to multiple voiceprint recognition models; PGD-UAP generates smaller perturbations, reducing auditory changes but at a slower generation speed and with possible limitations imposed by the input space; C&W-UAP is suitable for attacking specific models and has the capability to generate adversarial samples for multiple classes; MSCW-UAP has a faster generation speed and lower computational complexity than C&W-UAP, but it is more susceptible to limitations imposed by the input space.

The UrbanSound8k dataset employs down sampling at 16 kHz to train and evaluate the model while generating generalizations against perturbations. The dataset comprises 7.3 h of recordings subdivided into 8732 three-second long audio segments, representing each audio sample as a 50,999 dimensional array. The audio recordings were divided into ten distinct classifications, specifically, air_conditioner (sound of air conditioning), car_horn

(sound of a vehicle horn), children_playing (sound of children at play), dog_bark (sound of a dog barking), drilling (sound of drilling), engine_idling (sound of a car's engine running), gun_shot, jackhammer, siren, and street_music. The training set encompasses 80% of the dataset, while the remaining 20% is employed as a test set. To create opposing disturbances, the 6984 samples in the training set were randomized with a small batch size of 100 samples for the greedy iteration-based method. The perturbations were assessed for the entire test set, consisting of 1748 samples.

6.1. Evaluation Indicators

(1) Attack Success Rate

The attack success rate (ASR) is the likelihood of an adversarial perturbation added to an original sample causing the target classification model to make a classification error. ASR is calculated differently for the two distinct attacks, untargeted attack and target attack. If an adversarial sample with added perturbation is input into a target classification model during an untargeted attack and the model classifies the sample as any label other than the true label with high confidence, the attack is considered successful. The attack success rate under an untargeted attack is defined in Equation (8), where y_i represents the true label, and $f(\cdot)$ is the voiceprint classification model.

$$ASR = \frac{\sum_i f(x_i) \neq y_i}{\sum_i x_i} \tag{8}$$

In a targeted attack, given the attack target label t_i, select any sample and its true label $y_i \neq t_i$. The target attack succeeds if, after input to the target model, it is classified with high confidence as the target label t_i. The definition of the attack success rate under a targeted attack is presented in Equation (9).

$$ASR = \frac{\sum_i f(x_i) = y_i}{\sum_i x_i} \tag{9}$$

(2) Mean Generation Time for Adversarial Samples

The average generation time of the adversarial samples is calculated by dividing the overall generation time T of the antagonistic samples by the total number of antagonistic samples N, as expressed in Equation (10).

$$t = \frac{T}{N} \tag{10}$$

(3) Signal-to-Noise Ratio of the Counter Sample

In the audio domain, the paradigm constraint and signal-to-noise ratio are used to measure the anti-perturbation performance in a comprehensive way, and the SNR is calculated as shown in Equation (11), where P(v) denotes the root mean square (RMS) of the anti-perturbation signal v. The SNR of the anti-perturbation signal v is calculated as shown in Equation (11). All experiments in this paper take the average SNR of all adversarial samples as the evaluation index.

$$SNR = 10 \log_{10} P(v) \tag{11}$$

(4) Loudness of the Adversarial Sample

A high signal-to-noise ratio implies a low level of added noise through the generic counterattack. Moreover, the Celsius decibel (dB) of the adversarial perturbation, which measures the loudness of the samples, serves as one of the performance metrics. It is calculated as depicted in Equation (12). This metric bears some similarity to the $\boldsymbol{l}_\infty$ paradigm in the image domain, where a lower value conveys that the adversarial samples are only slightly distorted compared to the genuine audio samples. Lately, researchers have em-

ployed this technique to gauge the effectiveness of adversarial attacks on voiceprint classification models.

$$l_{dB_x}(v) = l_{dB}(v) - l_{dB}(x) \tag{12}$$

where $l_{dB}(x)$ is calculated as shown in Equation (13).

$$l_{dB}(x) = max_n(20\log_{10}(x_n)) \tag{13}$$

6.2. Untargeted Attack

In the absence of a specific target, the objective of a malicious attacker is to deliberately cause misclassification of input samples by the classification model, resulting in a confidence level that deviates significantly from the actual label. Hence, in this section, we calculate the optimization problem of the objective function via Equation (7), derive the attack success rate via Equation (8), determine the average signal-to-noise ratio via Equation (11), and evaluate the loudness utilizing Equation (12) during experimental analysis of data.

This section compares the effectiveness of generating adversarial samples for the five voiceprint classification models (1DCNN Rand, 1DCNN Gamma, ENVnet-V2, Sincnet, and SincNet+VGG19) in an untargeted attack scenario using the UrbanSound8K dataset. The experimental hyperparameters can be found in Table 3.

Table 3. Hyperparameters of experiments.

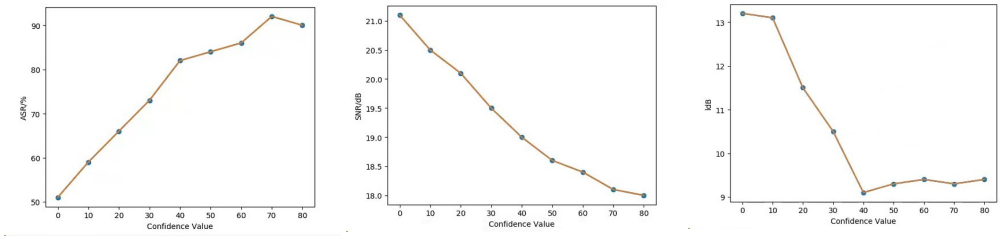
Parameter Type	Parameter Value
Confidence κ	40
Expected Deception Rate δ	0.1
Number of Iterations Epoch	100
Penalty Factor c	0.2

Due to the potential impact of different confidence levels (κ) on experimental results, the ENVnet-V2 and 1DCNN Gamma models were chosen as experimental objects to explore their impact on ASR, SNR, and l_{dB}. The confidence interval used was [0, 80], with confidence levels incremented by 10 at a time. The experimental results for the ENVnet-V2 model are displayed in Figure 2a–c. The experimental results for the 1DCNN Gamma model are shown in Figure 3a–c.

From Figures 2 and 3, it is evident that the ASR displays a general upward trend as the confidence level improves, while the SNR exhibits an overall downward trend with improvement in confidence levels. Additionally, the l_{dB} shows a considerable decrease in the range of [0, 40], followed by a leveling off in the range of [40, 80] with an increase in the confidence level. Despite being two distinct models for acoustic pattern classification, these observations hold true. Therefore, when selecting the hyperparameter for the untargeted attack experiment, if a confidence level of 40 is chosen, the impact on the results of SNR, ASR, and l_{dB} can be considered in the comparison of different models. This allows for more stable and representative experimental result values. The corresponding experimental results are shown in Table 4.

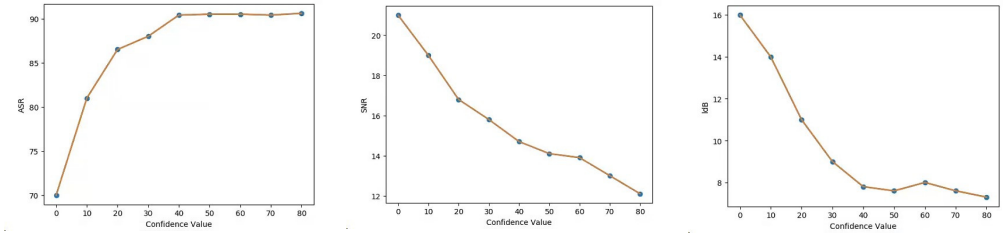
Table 4 shows that the success rate of the untargeted attacks on the five popular acoustic pattern classification models, with the help of acoustic pattern confrontation samples produced under the UrbanSound8K dataset, is the highest on the SincNet model from an ASR index perspective, reaching 90.4%, and the lowest is on the ENVnet-V2 model, but it also exhibits a superior attack level of 82.9%. The disparity between the training and test set ASR outcomes is minimal, with a maximum variance of only 6.2%. These findings suggest that the IG-UAP algorithm effectively curbs overfitting in untargeted attack situations. Based on SNR metrics, the SincNet model shows the highest SNR level of 29.886 dB, suggesting that adding the perturbation has minimal impact on the original acoustic pattern signal. The other four methods display comparable SNR levels with a lower effect on the initial signal. The ENVnet-V2, on the other hand, has the lowest SNR

level of 18.425 dB. From an l_{dB} metrics perspective, the SincNet model shows the smallest value, −26.214, while the ENVnet-V2 presents the highest value, −11.320. Both models maintain a low loudness that is not easily detected by the human ear.



(**a**) (**b**) (**c**)

Figure 2. Experimental results of ENVnet-V2 model at different confidence levels. (**a**) Effect of confidence value on ASR; (**b**) effect of confidence value on SNR; (**c**) effect of confidence value on l_dB.



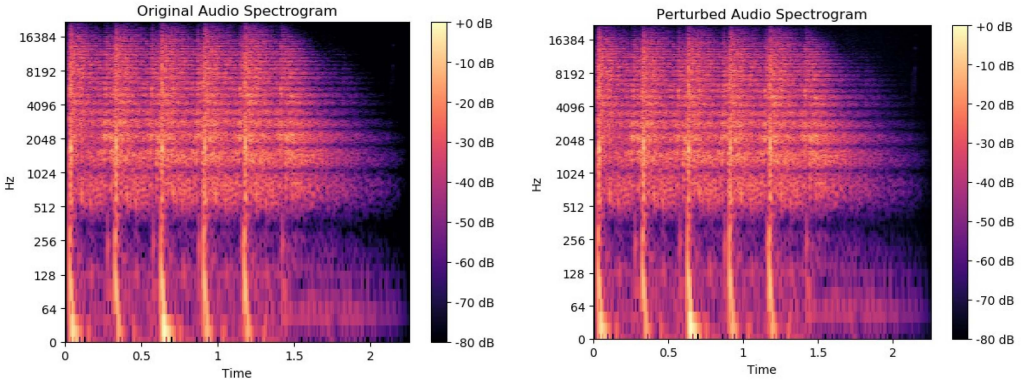
(**a**) (**b**) (**c**)

Figure 3. Experimental results of 1DCNN Gamma model at different confidence levels. (**a**) Effect of confidence value on ASR; (**b**) effect of confidence value on SNR; (**c**) effect of confidence value on l_dB.

Table 4. The success rate of IG-UAP on multiple target models.

Target Model	Datesets			
	Training Sets	Testing Sets		
	ASR/%	ASR/%	SNR/dB	l_{dB}
1DCNN Rand	89.2	86.5	20.168	−12.984
1DCNN Gamma	89.4	84.2	20.431	−18.451
ENVnet-V2	89.1	82.9	18.425	−11.320
SincNet	90.1	90.4	29.886	−26.214
SincNet+VGG19	88.3	85.2	23.346	−17.952

The results of the experiment are displayed in Figure 4. In this figure, we selected audio samples from the gun_shot category, to which we added the acoustic pattern generality adversarial perturbation. Figure 4a illustrates the original audio sample spectrogram, and Figure 4b portrays the spectrogram of the audio sample after adding the generality adversarial perturbation. This section bases its comparison experiments on the chosen 1DCNN Gamma target model and UrbanSound8K dataset, analyzing five approach methods in the no target attack context. These five methods include IG-UAP, FGSM-UAP, PGD-UAP, C&W-UAP, and MSCW-UAP, which are proposed in this chapter. The batch processing method utilizes the mean value, while some hyperparameters are prioritized during high performance. Table 5 presents the specific experimental hyperparameters, and the experimental results can be seen in Table 6.



(**a**) (**b**)

Figure 4. Effectiveness of five voiceprint classification models in the context of untargeted attacks. (**a**) Original spectrogram of the audio sample; (**b**) adversarial spectrogram of the audio sample.

Table 5. Hyperparameters of experiments.

Attack Methods	**Parameter Type**							
-	α	ε	**Learning Rate**	**Epoch**	**Batch Size**	**Confidence** κ	**Expected Spoofing Rate** δ	**Penalty Factor c**
IG-UAP	-	-	-	100	32	40	0.1	0.2
FGSM-UAP	-	0.1	-	100	32	-	-	-
PGD-UAP	0.1	0.0005	0.00001	100	32	-	-	-
C&W-UAP	0.1	-	0.0001	100	32	-	-	-
MSCW-UAP	1.5	-	0.001	100	32	-	-	-

Table 6. Comparative experimental results of untargeted attack.

Attack Methods	**Evaluation Indicators**		
	ASR/%	**Average SNR/dB**	**Average Spawn Time/s**
IG-UAP	**85.6**	20.551	20.168
FGSM-UAP	64.78	22.42	**0.55**
PGD-UAP	47.6	**39.05**	4.22
C&W-UAP	21.39	90.4	4.7
MSCW-UAP	30.12	24.89	14.66

After comparing the experiments, it has been observed that the method exhibiting the highest attack success rate is IG-UAP, achieving an ASR of 85.6%. Despite having the second-fastest average generation time, it exhibits the lowest average signal-to-noise ratio value, measuring at 20.551 dB. The fastest generation method is FGSM-UAP, clocking in with an average generation time of 0.55 s, which is notably higher than the other generation methods. The PGD-UAP method has the highest average SNR value of 39.05 dB, which is significantly higher than other generation methods. The IG-UAP method proposed in this chapter has the highest ASR and the best attack effect, while also having a longer average generation time compared to the PGD-UAP and C&W-UAP methods. However, the IG-UAP method has a higher production efficiency for the antagonistic samples. By combining all of the indicators, it can be concluded that the IG-UAP method is the most effective. Combined with the above findings, IG-UAP exhibits the most exceptional overall performance in creating acoustic generalizations against perturbations, compared to the other four methods, within the context of untargeted attacks.

In addressing the issue of varying sample sizes in different datasets, it may occur that some datasets have few samples available. To mitigate this, the IG-UAP algorithm

was employed for an experimental analysis of untargeted attacks on the 1DCNN Gamma model, whereby a single sample from each category in the UrbanSound8k dataset was taken. Table 3 shows the experimental hyperparameters, with the horizontal coordinates ranging from 0 to 9 in representation of the categories in the UrbanSound8k dataset. Figure 5 illustrates the experimental outcomes.

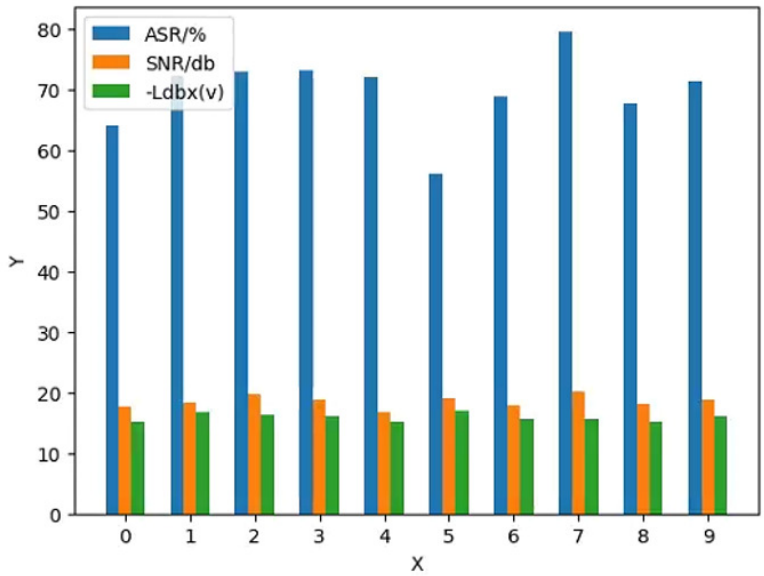


Figure 5. Attack effect diagram for a single sample.

The results of the experiment demonstrate that even when a single sample is taken from each category in the UrbanSound8k dataset, the IG-UAP method remains capable of computing effective acoustic generality against perturbations. For the small-sample size dataset, the average ASR of the acoustic generality anti-perturbation generated by the IG-UAP method is 68.8%, the average SNR is 19.125 dB, and the l_{dB} is −15.026 dB.

6.3. Targeted Attack

In targeted attacks, the objective is to cause the classification model to misclassify input samples with high confidence, resulting in the label specified by the attacker. In this section, we compute the optimization problem of the objective function using Equation (6), evaluate the attack success rate using Equation (8), calculate the average signal-to-noise ratio using Equation (11), and measure the loudness using Equation (12) for the purpose of data analysis and experiments.

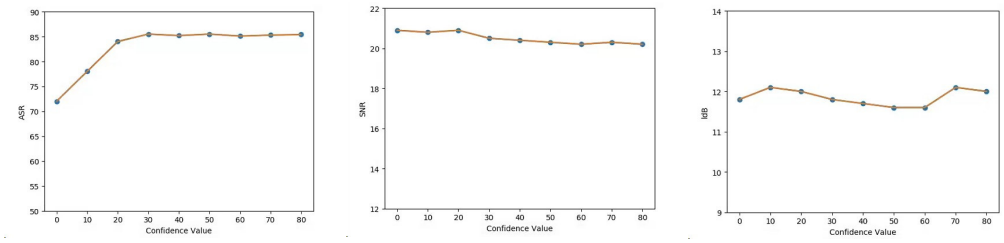
In this paper, we evaluate the efficacy of generating adversarial samples in the UrbanSound8K dataset for five voiceprint classification models, 1DCNN Rand, 1DCNN Gamma, ENVnet-V2, Sincnet, and SincNet+VGG19, in the presence of target-attack scenarios. The experimental hyperparameters are illustrated in Table 7.

Table 7. Hyperparameters of experiments.

Parameter Type	Parameter Value
Confidence κ	20
Expected Deception Rate δ	0.1
Number of Iterations Epoch	100
Penalty Factor c	0.15
Type of Targeted Attack	jackhammer

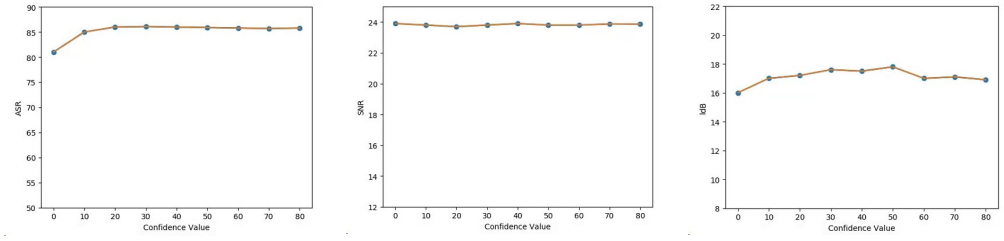
As different confidence levels can impact experimental outcomes, we chose the ENVnet-V2 and 1DCNN Gamma models as the objects of experimentation to evaluate their

effect on ASR, SNR, and l_{dB}. The experiments were conducted with a confidence interval of [0, 80], incrementing the confidence level by 10 for each test. The results for the ENVnet-V2 model are presented in Figure 6a–c, while the results for the 1DCNN Gamma model are presented in Figure 7a–c.



(**a**) (**b**) (**c**)

Figure 6. Experimental results of the ENVnet-V2 model at different confidence levels. (**a**) Effect of confidence value on ASR; (**b**) effect of confidence value on SNR; (**c**) effect of confidence value on l_dB.



(**a**) (**b**) (**c**)

Figure 7. Experimental results of the 1DCNN Gamma model at different confidence levels. (**a**) Effect of confidence value on ASR; (**b**) effect of confidence value on SNR; (**c**) effect of confidence value on l_dB.

Based on the data presented in Figures 6 and 7, two distinct acoustic pattern classification models were observed. The ASR exhibited an increase in the [0, 20] range and then plateaued between [20, 80] as the confidence level improved. The SNR showed a general trend of smoothness as the confidence level improved, whereas the l_dB also displayed an overall smooth trend with an increased confidence level. It is worth noting that the aforementioned results were observed despite the use of different classification models. Therefore, when selecting 20 as the confidence level hyperparameter in the target attack experiment carried out to compare various models, the effect of confidence level on the outcomes of SNR and ASR, as well as l_dB, can be factored in. Consequently, the experimental results are more stable and uniformly distributed.

The experimental results are shown in Table 8.

Table 8. The success rate of IG-UAP on multiple target models.

Target Model	Datesets			
	Training Sets	Testing Sets		
	ASR/%	ASR/%	SNR/dB	l_{dB}
1DCNN Rand	91.2	81.2	24.218	−17.043
1DCNN Gamma	90.9	83.7	22.654	−17.211
ENVnet-V2	91.9	83.1	21.146	−13.997
SincNet	96.5	90.9	30.241	−30.126
SincNet+VGG19	91.1	85.8	26.667	−20.912

According to Table 8, following an untargeted attack on five popular acoustic classification models, namely 1DCNN Rand, 1DCNN Gamma, ENVnet-V2, SincNet, and

SincNet+VGG19, using generated acoustic confrontation samples from the UrbanSound8K dataset, and assessed using ASR metrics, the SincNet model achieved the highest attack success rate of 90%. The attack success rate on the 1DCNN Rand model is the lowest, at 9%. However, it also demonstrates superior attack levels of 81.2%. The difference between the ASR outcomes on the training and test sets remains low, with a maximum differential of only 10.0%. This suggests that the IG-UAP algorithm is somewhat effective in reducing overfitting occurrences during targeted attacks. Analyzing the SNR metrics reveals that the SincNet model exhibits the highest SNR level, measuring at 30.241 dB. This indicates that the perturbation had a lesser impact on the acoustic pattern signal's original quality. Additionally, the other four methods exhibited similar SNR levels, which also maintained a lower impact on the original signal. ENVnet-V2, on the other hand, recorded the lowest SNR level at 21.146 dB. From the l_dB metrics' perspective, it was found that the SincNet model had the lowest value at -30.126, while the 1DCNN Rand model had the highest value at -17.043. The interval of loudness levels falls within a range that is difficult to detect with the human ear.

The paper findings are presented in Figure 8, where the engine_idling sound category samples were selected and subjected to generic adversarial perturbation. Figure 8a displays the spectrogram of the original audio sample, while Figure 8b exhibits the spectrogram of the audio sample following the generic adversarial perturbation.

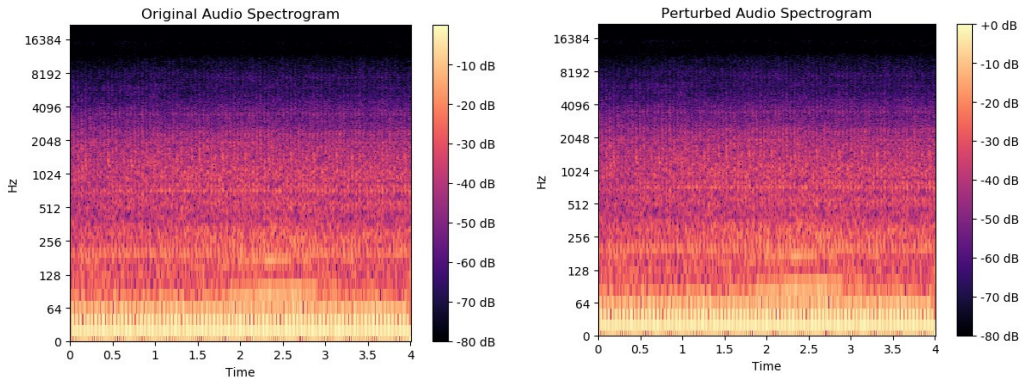


Figure 8. Comparison of original and added adversarial perturbation audio sample spectrograms. (**a**) Original spectrogram of the audio sample; (**b**) adversarial spectrogram of the audio sample.

After carrying out attacks on various models and analyzing the experimental results, we proceed with a comparison experiment using similar methods. It should be noted that the FGSM-UAP method and the PGD-UAP method typically generate untargeted adversary samples, which means that there is no guarantee that the resulting adversary samples will be classified into the specified target category, even if the attacker specifies the category. Therefore, this comparison of three methods, VCGAN-UAP, C&W-UAP, and MSCW-UAP, in the context of a target attack is presented. SincNet was chosen as the target model and UrbanSound8K as the dataset, with the target attack category being dog_dark. The experimental hyperparameters are listed in Table 9. It is important to note that this analysis aims to achieve objectivity, coherence, and clarity, maintaining an appropriate academic structure and language register whilst avoiding biased language, complex terminology, and subjective evaluations. The experimental results are shown in Table 10.

After conducting comparative experiments, it was found that when targeting the "dog_dark" category, the IG-UAP method demonstrates a superior attack success rate, an average signal-to-noise ratio (ASR), and an average generation time. The ASR reached 86.7%, which is 2.2% higher than the MSCW-UAP approach. Additionally, the average generation time of the IG-UAP method was 8.28 s faster than that of the MSCW-UAP approach, resulting in a significantly improved generation speed without compromising the attack success rate. The mean SNR is 28.145 dB, surpassing the C&W-UAP approach by

0.31 dB, yet the assault's success rate is 46.5% better, which guarantees a high SNR with a high accuracy level. Additionally, the typical generation time is 2.96 s. From a comprehensive analysis of each index, the IG-UAP method proposed in this chapter demonstrates a superior attack effectiveness and is the most efficient technique for generating voiceprints compared to the other two iterative methods in the context of a targeted attack. The method also exhibits excellent generalizability against perturbations and delivers the best overall performance.

Table 9. Hyperparameters of experiments.

Attack Methods	Parameter Type							
-	α	ε	Learning Rate	Epoch	Batch Size	Confidence κ	Expected Spoofing Rate δ	Penalty Factor c
IG-UAP	-	-	0.0002	100	32	20	0.1	0.15
C&W-UAP	0.1	-	0.0001	100	32	-	-	-
MSCW-UAP	1.5	-	0.001	100	32	-	-	-

Table 10. The success rate of IG-UAP on multiple target models.

Attack Methods	Evaluation Indicators		
	ASR/%	Average SNR/dB	Average Spawn Time/s
IG-UAP	86.7	28.145	2.96
PGD-UAP	40.2	27.835	4.22
MSCW-UAP	84.5	21.271	11.24

For the problem of large differences in the number of samples in different datasets, there may be cases where the number of samples in the dataset is small, therefore, under the condition of taking a single sample for each category in the UrbanSound8k dataset, the IG-UAP algorithm was used to test and experimentally analyze the target attack on the 1DCNN Gamma model, and the experimental hyperparameters were as shown in Table 6, in which horizontal coordinates 0 to 9 represent the categories in the UranSound8k dataset, respectively, as follows: air_conditioner, car_horn, children_playing, dog_bark, drilling, engine_idling, gun_shot, jackhammer, siren, and street_music. The ASR mixing matrix of the experimental results is shown in Figure 9.

When selecting jackhammer as the target category, the effectiveness of the targeting model is closely tied to the level of brightness in the mixing matrix. The darker the color, the greater the likelihood of successful classification into the target category, resulting in a better target attack. Conversely, the lighter the color, the lesser the chance of classification, indicating a less effective target attack. It is important to note that technical term abbreviations should be explained when first used. Regular author and institution formatting should also be maintained. According to the mixing matrix in Figure 9, the column dedicated to the jackhammer category appears darker in color when compared to the other categories, and it shows the highest probability for being classified as jackhammer during prediction, with a minimum likelihood of 82.4% and a maximum of 90.2%. Therefore, it can be demonstrated that the IG-UAP targeting performance is reliable and effective, even in a small dataset sample size.

6.4. Summary of the Experiment

In this chapter, evaluation metrics comprise the attack success rate, average signal-to-noise ratio, and loudness of the adversarial samples. Comparative experiments are conducted using the IG-UAP method to analyze five different acoustic pattern classification models under untargeted-attack and targeted-attack scenarios. In the untargeted attack, the five voiceprint classification models achieve an average attack success rate of 85.8%, with the lowest and highest values being 82.9% and 90.4%, respectively. The average signal-to-noise ratio is 22.451 dB, ranging from 18.425 dB to 29.886 dB, while the average

loudness is −17.384, with the lowest and highest values being −26.214 and −11.320. For jackhammer as the targeted attack category, the average success rate of the five acoustic classification models was 84.9%, with a minimum rate of 81.2% and a maximum of 90.9%. The average signal-to-noise ratio was 24.985 dB, with a range of 21.146 dB to 30.241 dB, and the average loudness was 19.858, with a minimum of −30.126 and a maximum of −13.997. Among the evaluation indices, the Generic Sound Generated Against Perturbation can be attacked effectively, and it exhibits consistent performance with minimal data variation.

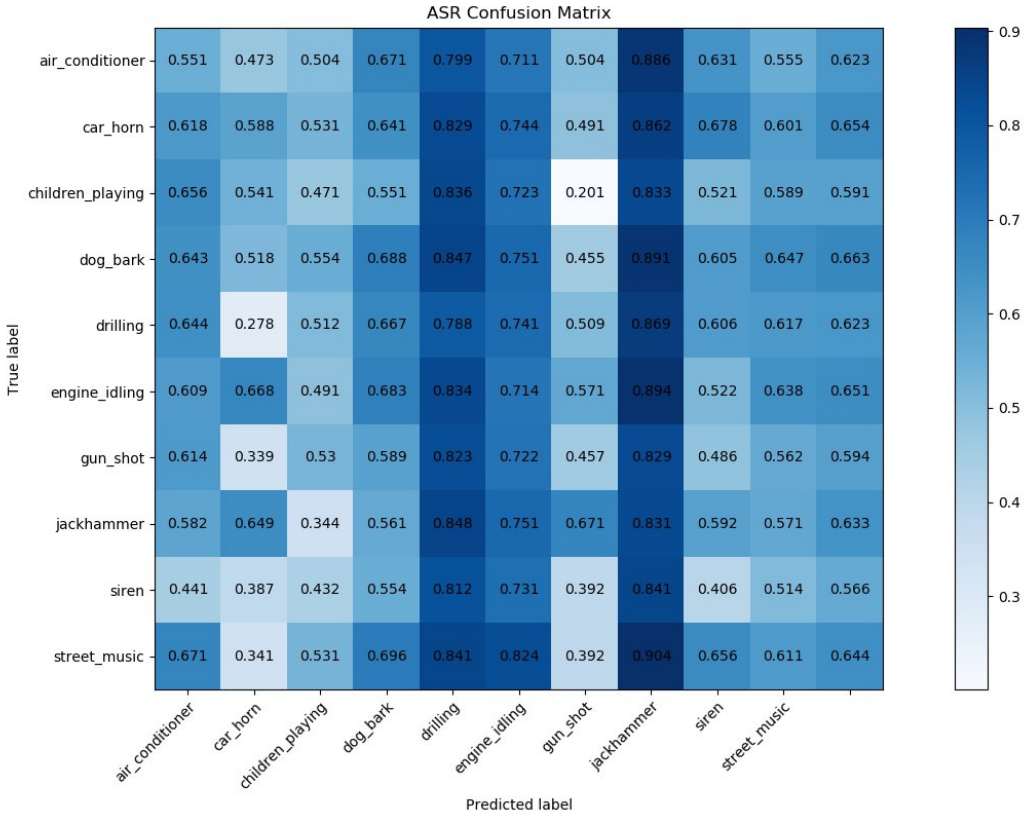


Figure 9. ASR confusion matrix.

In this paper, the evaluation indexes include the attack success rate, average signal-to-noise ratio, and average generation time of the adversarial samples. Using the IG-UAP method, we compare and analyze experimentally five similar acoustic UAP generation methods in both untargeted-attack and targeted-attack scenarios. Additionally, technical abbreviations are explained upon first use. In an untargeted attack, the IG-UAP achieves an 85.6% success rate, significantly surpassing the results of the four other generation methods. Its average generation time is 2.47 s, with only a 2 s difference from the fastest FGSM-UAP method. Additionally, the IG-UAP produces an average SNR of 20.551 dB, making it a highly effective tool for auditory masking. In the given targeted attack aimed at the dog_dark category, the IG-UAP method exhibits optimal evaluation indices. The attack success rate reaches 86.7%, and the average SNR is 28.145 dB. Additionally, the average generation time stands at 2.96 s. Conclusively, the IG-UAP method provides better acoustic generalization when dealing with perturbations as compared to other similar generation methods.

7. Conclusions

In this paper, the IG-UAP based method is elaborated in detail, including the symbolic description and detailed algorithmic flow of the IG-UAP method. The IG-UAP method guides the generation process of generic acoustic antiperturbation by using an influence function to attack the acoustic target classification model. To verify the method's effective-

ness, we conducted a series of experimental tests and analyses. Firstly, we compared it with similar methods in two cases: target attack and no target attack. The experimental results showed that the acoustic genericity antiperturbation generated using the IG-UAP method has a significant advantage in terms of attack effectiveness. Additionally, the experiments involved selecting and testing five distinct acoustic target classification models. The results of the experiments demonstrate that the acoustic generality adversarial perturbation generated by the IG-UAP method can effectively attack these diverse models, thus confirming the method's effectiveness and robustness. Based on these experimental results, it can be concluded that the IG-UAP method is an effective acoustic pattern adversarial attack method. It can generate acoustic pattern adversarial perturbations that deceive acoustic pattern target classification models with generality.

Author Contributions: Methodology, M.B. and J.X.; validation, M.B. and X.Y.; writing—original draft preparation, M.B. and J.X.; writing—review and editing, X.Y. and Z.J.; visualization, M.B. All authors have read and agreed to the published version of the manuscript

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets can be found in the following repositories: https://www.kaggle.com/datasets/chrisfilo/urbansound8k, accessed on 18 December 2023, reference [10].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity mappings in deep residual networks. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part IV 14; Springer: Cham Switzerland, 2016; pp. 630–645.
2. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
3. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegasm NV, USA, 26 June–1 July 2016; pp. 770–778.
4. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
5. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing properties of neural networks. *arXiv* **2013**, arXiv:1312.6199.
6. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, 7–12 June 2015; pp. 1–9.
7. Moosavi-Dezfooli, S.M.; Fawzi, A.; Fawzi, O.; Frossard, P. Universal adversarial perturbations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1765–1773.
8. Moosavi-Dezfooli, S.M.; Fawzi, A.; Fawzi, O.; Frossard, P.; Soatto, S. Robustness of classifiers to universal perturbations: A geometric perspective. *arXiv* **2017**, arXiv:1705.09554.
9. Neekhara, P.; Hussain, S.; Pandey, P.; Dubnov, S.; McAuley, J.; Koushanfar, F. Universal adversarial perturbations for speech recognition systems. *arXiv* **2019**, arXiv:1905.03828.
10. Salamon, J.; Jacoby, C.; Bello, J.P. A dataset and taxonomy for urban sound research. In Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; pp. 1041–1044.
11. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. *arXiv* **2014**, arXiv:1412.6572.
12. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv* **2017**, arXiv:1706.06083.
13. Carlini, N.; Wagner, D. Towards evaluating the robustness of neural networks. In Proceedings of the IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 22–26 May 2017; pp. 39–57.
14. Prinz, K.; Flexer, A. End-to-End Adversarial White Box Attacks on Music Instrument Classification. *arXiv* **2020**, arXiv:2007.14714.