A Comparative Study on Audio Adversarial Attacks

Omprakash Yadav[1], Angelica Sebastian[2], Melita Lewis[3], Sushree Nadiminty[4], and Shaun Noronha[5]
[1]*Assistant Professor, Department of Computer Engineering, Xavier Institute of Engineering, Mumbai, Maharashtra, India.*
[2,3,4,5]*Student, Department of Computer Engineering, Xavier Institute of Engineering, Mumbai, Maharashtra, India.*

***Abstract*— Adversarial examples are prone to neural networks when specific types of inputs are given to a system that can result in misclassification or incorrect output. With the growing prominence of personal voice assistants (Google Home, Siri, Alexa, etc.) which depend on Automatic Speech Recognition systems (ASR) which are an application of neural networks, a question arises as to how robust these systems are to adversarial attacks. This makes adversarial audio attacks a critical topic in the current world of automated systems. This paper aims at presenting a thorough introduction to the background knowledge of adversarial attacks, and the generation of adversarial examples as well as psychoacoustic models and the different evaluation indicators. It's necessary to understand how the Deep Learning models in Automatic Speech Recognition systems (ASR) are vulnerable to attacks and how these attacks are performed using different methods**

***Index Terms*— ASR, Attack, Audio Adversarial, Carlini, Comparison, Neural Network, Psychoacoustics.**

I. INTRODUCTION

Deep neural networks (DNNs), the most effective artificial intelligence (AI) method currently available, are widely used in a variety of real-world applications. Despite their current popularity and effectiveness, DNNs have many serious flaws, particularly a high sensitivity to adversarial attacks, an extremely detrimental attack strategy that introduces carefully designed adversarial perturbation given to the DNNs' benign input which can lead to misclassification. While comparing the existing systems, we realized that little research has been done on audio adversarial examples against speech recognition models as compared to image adversarial examples against image classification models. Deep neural network systems have been demonstrated to be open to hostile attacks. This presents a chance for lawbreakers and poses a threat to the protection of personal information and property. The safety of the public's privacy is in jeopardy from a security standpoint. New varieties of speech systems could continually arise as science and technology advance, but the vulnerability of neural networks to attackers remains an issue. Therefore, it is crucial and essential to conduct an overview of existing technology before tackling new problems. Therefore, mastering attack strategies will help us to stop issues before their likely appearance, thereby promoting both personal and public safety. Although such models are widely used in various commercial applications like Amazon Alexa, Apple Siri, Microsoft Cortana, Google Assistant, and other home automated devices like Amazon Echo and Google Home, little is known about the implications of an adversarial attack on these applications. Therefore we aim to study how an audio adversarial attack takes place and what are the different methods that are used to perform these adversarial attacks.

II. BACKGROUND

A. Deep Learning

Deep learning is a small subset of a larger family of machine learning methods, where 'deep' refers to the use of the multiple layers in the network. To extract higher-level features from raw data, deep learning algorithms employ multiple layers. Consider image processing, where lower layers may identify as edges and higher layers may identify as concepts significant to people, such as figures, texts, or faces[1]. One of the most significant advantages of deep learning is its ability to work with unstructured data. Because the majority of business data is unstructured, such as text, photos, and speech, deep learning is a helpful tool. The various layers of deep neural networks allow models to become more efficient at learning complex information and performing increasingly complex

computational tasks. It also outperforms machine learning algorithms in unstructured dataset machine perception tasks (the capacity to sort through inputs like images, audio, and video like a human). Deep learning has greatly increased accuracy, particularly in image classification and speech recognition, and is now widely employed.

B. Deep Learning Models

- Recurrent Neural Networks - RNNs RNNs are neural networks that allow data to flow in either direction. The fundamental idea behind RNNs is to leverage sequential information. The primary assumption of a normal neural network is that all inputs and outputs are independent of one another. For example, if we want to predict the next word in a string of words, we must first know which words came before it.
- Convolutional Deep Neural Networks - CNNs CNNs are mostly employed in computer vision, but they are also utilised in acoustic modelling for automatic speech recognition. Convolutional neural networks are based on the concept of a moving filter that goes across an image. This moving filter, also known as convolution, is applied to a specific region of nodes. Using pixels in an image as an example, the filter used may be something like 0.5 x the node value.
- Generative Adversarial Networks - GANs Generative adversarial networks are a type of deep neural network made up of two nets that compete with each other. The generator neural network generates new data instances, whereas the discriminator neural network examines these instances for authenticity. Both of these networks learn from one another and thus grow. Consider the generation of hand-written numerals from the actual world, such as those seen in the MNIST dataset. The discriminator's job is to identify an instance from the real MNIST dataset as authentic.

C. Adversarial Attacks

Deep learning methods are known for being sensitive to adversarial examples. This means that an attacker can purposefully alter the examples to cause a given model to misclassify a particular input. Such examples are known as adversarial examples[2].The first adversarial examples focused on image recognition systems, but they were eventually expanded to include speech recognition, speaker recognition, and other systems.

D.Audio Adversarial Attacks

While the first adversarial examples were in the image domain, later attacks on ASR systems, as discussed in the following sections, have demonstrated the existence of adversarial examples in the audio domain. The addition of perturbation to the original signal allows the original audio signal to be transcribed to a target phrase requested by the adversary or causes significant transcription error by the victim ASR model.

E. Psychoacoustics

The study of how humans detect sound with their ears and what they experience, as well as the statistical relationships between acoustic stimuli and hearing experiences, is known as psychoacoustics. The frequency, intensity, and interference from other sounds influence the human ear's capacity to perceive a sound signal. The crucial aspect of a provided audio adversarial example is that it should not be detected by human hearing. The psychoacoustic model seeks to conceal noise in the audio transmission so that it is imperceptible to human ears.

III. RELATED WORK

Nicholas Carlini and his adviser, David Wagner, developed audio adversarial examples for use with automatic speech recognition systems. Their main goal was to be able to make another audio waveform that was more than 99.9% similar to any given audio waveform, but the transcription of this newly produced audio waveform may be any word that we choose. On Mozilla's DeepSpeech implementation, it was a white-box iterative optimization-based attack. It was a complete attack with a 100% success rate.[3].

Hiromu Yakura and Jun Sakuma suggested a way for creating an audio adversarial example that may be used to challenge a state-of-the-art speech recognition machine in the real world. It generates robust adversarial instances by physically simulating the alterations generated by playback or recording. It was an attack that was completely unnoticeable by us humans[2].

Joseph Szurley and J.Zico Kotler suggested a method for generating automated room impulse responses based on a psychoacoustic-property-based loss function. Their goal was to develop an adversarial attack that could withstand being broadcasted through a speaker in one or more environments. They showed that such attacks were possible even when it was virtually imperceptible to the listeners[4].

Das Nilaksh et al. (2018) developed a tool that allowed real-time interactive testing with adversarial assaults and defences on an automatic speech recognition system. It was the first interactive tool that could experiment both graphically and audibly. "ADAGIO" was the name given to this tool. It included Adaptive Multi-Rate and MP3 audio compression algorithms as defences that users may apply interactively to the assaulted audio samples. These strategies were based on psychoacoustic ideas, which successfully eliminated targeted attacks, decreasing the attack success rate from 92.5% to 0%[5].

IV. REVIEW ON DIFFERENT THREAT MODELS

A. Audio Adversarial Attack using Generative model

Deep Neural Networks (DNNs)-based applications that take audio as input are vulnerable to possible adversarial attacks. Although the existing methods to perform audio adversarial attacks are successful, there are quite a few challenges that come along the way. One of them is the large time budget that is required to generate an adversarial perturbation. Most audio-domain applications take quick streaming inputs and process them in real-time. In such cases, these time-consuming constraints slow down the process of launching an attack on the system [6].

Another challenge is the observation of the full content of the audio input. Since it is difficult to know the entire content of the ongoing audio input throughout its input streaming phase, the majority of adversarial audio-generating algorithms are extremely unrealistic when used in real-time audio applications [6]. We may therefore use a generative model to generate adversarial perturbations for audio input in a single forward pass in order to overcome these limitations. Firstly, a generative model needs to be trained by distributing the data itself into training data and testing data.

Once it is well-trained, it can generate adversarial perturbations quite quickly. Thus, launching an adversarial attack on the audio-domain systems is possible by using a generative model that can accelerate the speed of perturbation generation in a real-time setting. Several image-domain adversarial attacks generated image adversarial perturbations using traditional generative models like the Generative Adversarial Network (GAN) and autoencoder. But in image-domain-based adversarial attacks, different models were required for different target classes. However, for audio-domain-based adversarial attacks, we can use a single generative model for any adversary's desired class. Both Targeted and Untargeted adversarial attacks can be performed using generative models. We can generate input-dependent as well input-independent perturbations with the help of generative models. In real-life scenarios, it is quite unlikely to know the contents of the audio input beforehand as the audio signals in the input have an inherent temporal sequence.

As a result, input-independent audio adversarial attacks are more feasible because they don't require the need to observe the entire audio input's content. To make sure that a single generative model may be applied again to analyse audio input from a new target class, the generative model makes use of embedding feature maps.

Therefore, during the training phase, embedding feature maps are jointly trained with the training dataset. Once it has been trained properly, we can give any audio input and a target class label and it will generate an adversarial perturbation by performing inference on the audio input. The generative model's intermediate feature map is then concatenated with the embedding feature map for the target class [6].

B. Audio Adversarial Attack using Particle Swarm Optimization (PSO) and Fooling Gradient Method

End-to-End Acoustic systems are systems that translate a series of acoustic features from an input signal into a sequence of distinct letters or words. These systems are categorized as recognition-oriented systems and classification oriented systems [8]. In recognition-oriented systems, the audio input is first divided into frames and then the corresponding output of each of these frames is predicted. It then derives the recognized output using Connectionist-

TemporalClassification (CTC) loss or other techniques.

In contrast, classification-oriented systems carry out classification, a type of supervised learning, on the corresponding spectrograms after the audio input is first translated from the time domain into the frequency domain. Yet, the majority of acoustic systems are constructed using deep neural network models in order to get greater performance.

These models are vulnerable to adversarial audio attacks. A simple audio input can be manipulated in a malicious manner by adding adversarial audio into the actual audio in such a way that a human can hear the non-malicious command but the acoustic system will transcribe into a malicious command.

These attacks can be done using the Particle Swarm Optimization (PSO) algorithm and the fooling gradient method. PSO is a heuristic approach that is highly influenced by the behaviour of a swarm of birds in order to find the optimal solution. Without requiring any gradient information, this approach can be used to search a huge area of potential optimal solutions. Many alternative solutions are iteratively moved around in the search space according to how well they are suited to tackle a certain issue.

The fooling gradient method is used to compute the gradient of the input audio signals instead of the model parameters. This method is then used to disguise the malicious audio into the original audio input. This process is done iteratively. SIRENATTACK is a type of attack that uses PSO and the fooling gradient method to perform audio adversarial attack [8].

C. Universal Adversarial Audio Perturbations

In both targeted and untargeted attack scenarios, this paper illustrates the existence of universal adversarial perturbations that can deceive a range of audio classification architectures. It suggests two approaches for locating these perturbations. The first strategy, which is iterative and greedy, aggregates minor input perturbations to push them towards the decision boundary. The second technique involves a new penalty formulation that identifies both targeted and untargeted global adversarial perturbations. An suitable objective function is minimised on a batch of samples using the penalty method that has been presented. Additionally, it offers evidence that the suggested penalty technique theoretically converges to a result that relates to all adversarial perturbations. This study asserts that even with a single sample from the target dataset, it is still viable to conduct effective attacks utilising the penalty technique. Using the suggested penalty technique, the attack success percentages for targeted and untargeted attacks are also higher than 85.0 and 83.1 percent, respectively.

Finding a vector v that, when added to the audio samples, can trick the classifier into classifying the majority of the samples is the objective here. This vector, which is known as universal, can be introduced to any sample to deceive a classifier because it is a fixed perturbation that is independent of the audio samples[9].

D. Detection of adversarial attacks using deep learning techniques

This work implements an efficient detection method that establishes a temporal relationship between various AV streams. Deep Convolutional Neural Network (DCNN) technology is used for this. Two audio-visual recognition models that have been trained using Lip reading datasets are used in the proposed technique to detect adversarial attacks. These two models are the Geospatial Repository and Data (GRiD) Management model and Lip-Reading in the Wild (LRW). The proposed strategy is an effective way to recognise adversarial attacks when compared to Supervised Kernel Machines, Combined Neural Networks, and Band Feature Selection methods. Due to the potential for significant misclassification caused by adversarial instances, which are produced by adding only a small amount of sound to the original sample, deep learning approaches are vulnerable to such adversarial attacks. Three convolutional layers are used in this method, and the sliding window area is very small. The first and second convolutional layers employ the 32-layer linear Scaled Exponential Linear Unit (SELU) activation function. In contrast, the third one contains a 64-layer Rectified Linear Unit (ReLU) activation function. A maximum pooling layer comes after each convolutional layer. After that, the data is sent through a layer of the winding/pooling mechanism that is entirely connected, where the SoftMax layer calculates the binary performance of regular and malicious samples.

The MFCC method divides the signal into small frames by applying a Mel frequency filter bank to that specific frame and adding discrete cosine changes and logarithmic weights from the input signals. The

audiovisual features are extracted using this technique. Following feature extraction, the DCNN evaluates the feature values and determines if an adversarial approach is feasible or not. This research uses the GRID (Cooke et al. 2006) and LRW (Chung Zisserman, 2016) datasets to find malicious assaults.

E. Defense by reverse-engineering the added noise to cancel out the malicious noise

This work offers a solid defence against audio adversarial examples that are undetectable or inaudible. This method involves adding targeted proportional additive Gaussian noise to an adversarial example to restore it to its original transcription.

The first randomised or probabilistic defence, it performs comparably to other defences. It also demonstrates the difficulties encountered when attempting to implement defences against hostile examples for audio.

These certified defences for adversarial examples on images to defences for AAEs are challenging to adapt, generalise, or expand. This problem arises from the fact that all currently recognised defences exploit the adversarial tactic of introducing uniform low magnitude adversarial perturbations, which is used by many vision assaults.

However, this adversarial tactic is trivially distinct from others since an adversary might benefit from psychoacoustics, the academic study of how people perceive sound, while producing AAEs. In order to create imperceptible audio adversarial examples, the adversary can use psychoacoustics to introduce hostile noise into the precise audio regions that are inaudible to humans. Therefore, by introducing localised, high magnitude adversarial perturbations, invisible AAEs are produced, in contrast to the adversarial strategy that is neutralised by certified defences.

By taking into consideration this distinct adversary tactic and taking into account the particular auditory areas that these attacks target, this research develops an efficient defence specifically for undetectable AAEs

F. WaveGuard - Detection and defense against different adversarial attacks

WaveGuard, using different audio transformation functions, analyses the ASR transcriptions of original and changed sunds, so it can identify the adversarial examples. With the research done, the defence system was abe to consistently identify adversarial examples created by four different methods of adversarial attacks. They examined five distinct audio transformation functions under varying levels of compression so as to create a sample which could oppose non-adaptive systems. Linear Predictive Coding (LPC) and Mel spectrogram extraction-inversion are two new audio transformation functions that this work proposes that are more resistant to adaptive attacks than previous transformation functions.

But one limitation of the research done was that the attacks could not be performed in real time as each point in the audio that the attacker wishes to mistranscribe, requires the solution of a complex optimization problem. The authors created an algorithm to locate a single quasi-imperceptible universal perturbation that results in mistranscription when any random speech signal is supplied to the victim speech recognition model in order to carry out attacks in real time. A mistake in transcription by a speech recognition model can be produced by adding a universal perturbation vector to any speech waveform using the suggested approach, which repeatedly runs across the training dataset.

G. Audio Adversarial Attack using different Evaluation Indicators

An automatic speech recognition (ASR) system that is built upon a deep neural network can be attacked by producing incorrect transcriptions to a given audio input which is imperceptible to humans. In 2014, Limited memory-BFGS modelled as a constrained minimization problem was the first adversarial example that was introduced against a deep neural network. It uses a loss function such as cross entropy to minimize the distance in order to solve and get multiple values for the optimization problem. Although this algorithm has a quick generating time and a small memory footprint, its confrontation requires work.

The fast gradient sign method (FGSM) differs significantly from the L-BFGS approach in two important ways: first, it optimises for Linf ty distance measurement, and second, its main objective is to produce adversarial examples as fast as possible. FGSM has a lower success rate with a nonlinear model but is simpler and computationally more efficient than other methods. In a non-linear model, numerous

iterations are required to identify the ideal circumstance because, if just one iteration is carried out, the direction may not be entirely accurate.

BIM, or the basic iterative method, is employed here. The BIM approach divides a single FGSM step into numerous smaller ones in order to provide iterative adversarial examples. In a nonlinear model, this strategy can provide adversarial examples, albeit at a high computational cost. Another method is DeepFool, which uses the difference between the boundaries of the classifier and the input x to compute a minimal norm of adversarial perturbation. Iteration is used to establish the minimal standard to reduce the disturbance.

A technique for producing adversarial samples for the kind of deep neural network is called the Jacobian-based Saliency Map Attack (JSMA). To apply JSMA, forward guide numbers are utilised.

The genetic algorithm is a gradient-free optimization method that does not require any prior knowledge of the attacked systems. To generate a large number of adversarial example candidates, just add some random noise to a set of patterns in an audio clip. To reduce the impact of noise on humans, the sound must be placed in the least noticeable location of the random system of audio examples

V. CONCLUSION

In this paper, we briefly discuss how neural networks are susceptible to adversarial examples, which are specific inputs to a network that result in a misclassification or an incorrect output. An audio adversarial attack using a generative model which can perform a fast and universal audio adversarial attack on an automated speech recognition system is being studied along with an adversarial attack on deep neural network-based voice processing systems. Compared with existing attacks, SirenAttack is a new class of attacks to generate adversarial audios having significant features like stealth, versatility, and effectiveness. This paper also demonstrates the existence of universal adversarial perturbations, which can fool a family of audio classification architectures, for both targeted and untargeted attack scenarios. A framework called Waveguard is being used to detect the adversarial inputs crafted to attack ASR systems. Accordingly, audio input is divided into frames and converted into Mel-Frequency which is a perceptually relevant scale for pitch and using which MFCC(Mel-frequency cepstral coefficients) is calculated to perform feature extraction.

TABLE I
COMPARISON OF DIFFERENT THREAT MODELS.

Sr no.	Threat Model	Type of Attack	Adversarial Knowledge	Over the Air	Success Rate
1	Audio Adversarial Attacks using Generative Model	Targeted & Untargeted	White-box & Black-box	No	Targeted-96.35%, Untargeted-88.51%
2	Audio Adversarial Attack using Particle Swarm Optimization (PSO) and Fooling Gradient Method	Targeted	White-box & Black-box	Yes	99.45%
3	Universal Adversarial Audio Perturbations[9]	Targeted & Untargeted	White-box	No	Targeted-85.0%, Untargeted-83.1%
4	Detecting adversarial attacks on audio-visual speech recognition using deep learning method[10]	-	-	No	95.60%
5	Defending Against Imperceptible Audio Adversarial Examples Using Proportional Additive Gaussian Noise[11]	Targeted	White-box	Yes	-
6	WaveGuard: Understanding and Mitigating Audio Adversarial Examples[12]	Targeted & Untargeted	White-box	No	-
7	Audio Adversarial Attack using different Evaluation Indicators	-	White-box & Black-box	No	-

REFERENCES

[1] Deng, Li, and Dong Yu. ”Deep learning: methods and applications.” Foundations and trends® in signal processing 7.3–4 (2014): 197-387.
[2] Yakura, Hiromu, and Jun Sakuma. ”Robust audio adversarial example for a physical attack.” arXiv preprint arXiv:1810.11793 (2018).
[3] Carlini, Nicholas, and David Wagner. ”Audio adversarial examples: Targeted attacks on speech-to-text.” 2018 IEEE security and privacy workshops (SPW). IEEE, 2018.
[4] Szurley, Joseph, and J. Zico Kolter. ”Perceptual based adversarial audio attacks.” arXiv preprint arXiv:1906.06355 (2019).
[5] Das, Nilaksh, et al. ”Adagio: Interactive experimentation with adversarial attack and defense for audio.” Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Cham, 2018.
[6] Xie, Yi, et al. ”Enabling fast and universal audio adversarial attack using generative model.” Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35. No. 16. 2021.
[7] Chen, Xiaojiao, Sheng Li, and Hao Huang. ”Adversarial Attack and Defense on Deep Neural Network-Based Voice Processing Systems: An Overview.” Applied Sciences 11.18 (2021): 8450.
[8] Du, Tianyu, et al. ”Sirenattack: Generating adversarial audio for end-to-end acoustic systems.” Proceedings of the 15th ACM Asia Conference on Computer and Communications Security. 2020.
[9] Abdoli, Sajjad, et al. ”Universal adversarial audio perturbations.” arXiv preprint arXiv:1908.03173 (2019).
[10] Ramadan, Rabie A. ”Detecting adversarial attacks on audio-visual speech recognition using deep learning method.” International Journal of Speech Technology 25.3 (2022): 625-631.
[11] Mendes, Ethan, and Kyle Hogan. ”Defending against imperceptible audio adversarial examples using proportional additive gaussian noise.” (2020).
[12] Hussain, Shehzeen, et al. ”WaveGuard: Understanding and Mitigating Audio Adversarial Examples.” 30th USENIX Security Symposium (USENIX Security 21). 2021.