ROBUST NOISE-BASED ATTACKS AGAINST AUDIO EVENT DETECTION SYSTEMS

by

RODRIGO AUGUSTO SILVA DOS SANTOS

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy at
The University of Texas at Arlington
May, 2022

Arlington, Texas

Supervising Committee:

Shirin Nilizadeh, Supervising Professor
Bahram Khalili
Farhad Kamangar
Jiang Ming

ABSTRACT

Robust Noise-Based Attacks Against Audio Event Detection Systems

Rodrigo Augusto Silva dos Santos, Ph.D.

The University of Texas at Arlington, 2022

Supervising Professor: Shirin Nilizadeh

The massive advances on the field of deep neural networks in the 2000 and 2010 decades led to an overwhelming adoption of these algorithms on all sorts of domains and applications. Under this widespread adoption scenario, it is natural that these neural networks have also been employed on safety-related use cases, bringing substantial improvements to the performance of existing as well as novel systems. Examples of these safety-inclined applications include scene recognition, object detection and tracking, speech recognition, audio event detection and classification, just to cite a few ones.

Unfortunately, these neural network algorithms have been shown to be vulnerable to different forms of attacks that can prevent them from performing as intended and as designed. These attacks have also, so far, been shown to be impossible to be fully eliminated or even dealt with to a definitive degree of satisfaction. This is because these attacks exploit the very fundamental way these algorithms are conceived in the first place, deriving their malicious efficacy from the very intrinsic neural networks properties.

The focus of this dissertation is on audio event detection (AED) systems and on to seek to contribute for the advance of neural network safe use on the AED domain. Existing real AED systems are tested to exhaustion to evaluate the state-of-the-art. Research and implementation efforts are then switched to neural networks (NN), the main component behind the AED capabilities by several of these modern systems.

Throughout this doctoral research, different state-of-the-art AED devices are field tested, several AED classifiers are implemented, attacked, as well as defended, and a full End-to-end AED system is proposed. These experiments are done under the objective to generate new knowledge to contribute to the mitigation and bridging of the existing gaps in practical AED systems.

ACKNOWLEDGMENTS

I would like to thank my supervising Professor, Dr. Shirin Nilizadeh, for leading me towards the main technical path I took over the course of my studies, as well as for mentoring me throughout this difficult Ph.D. journey. I would like to thank my supervising committee for the patience and for the professionalism it had when working with me on my final contributions. I would like to thank Dr. Bahram Khalili, who besides being a mentor and a role model throughout the entire Ph.D. research, was an outstanding advisor that provided me with much needed support while I was progressing through the Ph.D. research steps.

DEDICATION

I would like to thank my beloved wife, Lorena Cavalcante, for all the support she gave me, as well the resilience she demonstrated during my Ph.D. journey.

LIST OF ILLUSTRATIONS

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	**Meaning**
ACC	Accuracy
AE	Adversarial Example
AED	Audio Event Detection
BN	Background Noise
CNN	Convolutional Neural Network
CPS	Cyber Physical System
CRNN	Convolutional Recurrent Neural Network
Db	Decibel
DCNN	Dilatated CNN
E2E	End-to-end
GMM	Gaussian Mixture Model
GPU	Graphics Processing Unit
GRU	Gated Recurrent Unit
IoT	Internet of Things
LSTM	Long Short-Term Memory
MFCC	Mel-frequency Cepstrum Coefficient
ML	Machine Learning
NN	Neural Network
PREC	Precision
RCL	Recall
ReLU	Rectified Linear Units
RNN	Recurrent Neural Network
SNR	Signal to Noise Ratio
SR	Speech Recognition
SVM	Support Vector Machine
WN	White Noise

TABLE OF CONTENTS

CHAPTER 1: AUDIO EVENT DETECTION

In this chapter, the foundations of this doctoral research are presented. As such, in section 1.1 the full research background as well as the two fundamental knowledge areas for this dissertation, namely deep neural networks and audio event detection are introduced; in section 1.2, an overview is provided on some of the relevant technical challenges pertaining the practical intersection of these two key knowledge areas; in section 1.3, an early overview is provided on the proposed scientific contributions to be originated from this doctoral research; in section 1.4 an outline of what will be addressed in each subsequent chapter of this dissertation is discussed, which includes an overview of the final work to be carried out during the last year of doctoral research.

1.1 RESEARCH BACKGROUND

The Internet of Things, or simply IoT, is a term introduced back in 1999 [Ahsan2016] that describes an unprecedented network made of electrical or heterogeneous “electronic devices of various sizes and capabilities that are connected to the internet” [Miraz2015]. According to [Larrucea2017], these “networked sensors and smart objects” serve the purpose of measuring / controlling / operating on an environment in order to make it intelligent, usable, programmable, and capable of providing useful services to humans".

Practical applications of IoT devices include but are not limited to environmental monitoring, infrastructure management, manufacturing, home

automation, smart cities, transportation, medical and health care systems, and others. According to [Hognelid2015], IoT has been described as the third wave of information technology-driven transformation, the first dating back to the 1960`s and 1970`s, when computers became able to perform previously manual tasks and processes, and the second dating back to the 1980s and 1990s, when connectivity and communication between machines and humans became ubiquitous.

According to [Husamuddin2017], by 2020, it is expected that IoT will reach 50 billion connected devices, and by 2025, according to [Al-Fuqaha2015], the whole annual economic impact to be caused by IoT will be over 6 trillion dollars. These forecasts, according to [Shrestha2014], offer great opportunity for IoT users to be connected to anything, whenever needed, wherever needed. With so many users' devices connected among themselves, one can infer that the amount of data collected by such devices is massive. To serve IoT users efficiently, these huge amounts of data should be worked in real-time [Al-Fuqaha2014].

In other words, the data must go through "several levels of processing in order to produce a high-level description of the environment with discrete semantic states called context" [Venkatesh2017]. This IoT vision, according to [Singh2014], will be built on top of the diverse sensors available to users, and these are the basic way in which these large volumes of raw data will be gathered for subsequent "churning out" in an understandable manner. Still within the presented context, IoT based Cyber-Physical Systems (CPSs) go even further, including not only embedded sensors and processors, but also actuators, allowing these systems not only to sense but also to interact with the physical world.

Many of these IoT CPS systems include audio-related capabilities.

1.1.1 Audio Event Detection

Audio-enabled CPS systems can obtain the audio input from some acoustic sensors and then can process it for some specific purpose. Among those purposes one could name speech recognition (SR) and audio event detection (AED), just to cite a few. While both SR and AED systems work on audio samples, their goals and algorithms are different. Speech recognition works on vocal tracts and structured language, where the units of sound (e.g., phonemes) are similar. As such, SR systems require an approach for linking these basic units of sounds to form words and sentences, which then in turn, become recognizable and meaningful to humans [Bilen2020, Cowling2003, Jose2020].

AED systems, on the other hand, cannot look for specific phonetic sequences to identify a specific sound [Hamid2014], and because of very distinct patterns presented by different sound events (e.g.: dog bark vs. gunshot) a different AED algorithm should be used for every specific sound event combo. Also, for AED, the Signal-To-Noise Ratio (SNR) tends to be low, being even lower when the distance between acoustic source and the microphones performing the audio capture increases [Crocco2016]. As such different authors indicate that developing algorithms for detecting audio events is more challenging than developing algorithms for SR [Bilen2020, Cowling2003, Hamid2014, Crocco2016].

Both SR and AED are relevant and applicable to a wide variety of domain problems. This doctoral dissertation, however, will focus on AED applications for the safety domain. This is due to safety being a major concern in people's lives. For instance, gun shooting represents one of the major threats to safety every

person is exposed to [Nytimes2018]. Situations like the Las Vegas Mandalay Bay hotel massacre, where a shooter fired his guns at defenseless and innocent country music concertgoers, killing 58 and harming over 850 people, are good examples of how everything can suddenly run out of control, bringing major impact to the lives of individuals, families, and authorities.

There have been over 300 mass shootings in the US in 2018 alone, according to the Business Insider [Robinson2018], which contributes to statistics pointing out to the trend that it is more likely that Americans will die to gun violence than many of the other leading causes of deaths combined. With numbers such as these, it is not surprising that governmental efforts and reports such as [HHS2014] try to convey to organizations and the general public the need for preparedness for situations that involve risks to safety, such as that of active shooters.

The approach on [HHS2014] as well as other similar approaches usually rely on some sort of planning and preventive actions, followed by response actions to be employed once the emergency occurs. While one cannot overstate how important such efforts are in improving the general public's response to emergency situations, this doctoral dissertation also advocates for the continuous advance on the field of emergency technology through research efforts that ultimately lead to new knowledge as well as technological improvements.

This work, thus seek to achieve such improvements through the promotion of advances on safety-related application of AED systems, capable of detection and subsequent classification of safety related sonic events of interest, such as gunshots and glass breakage. In the last decades there has been a surge on the research and development of Machine Learning-enabled AED systems, and this dissertation leverages this trend.

1.1.2 Deep Learning and Neural Networks

Machine Learning, while a subfield of Artificial Intelligence, is, according to [Somvanshi2016], “the ability of machines to learn”, through the employment of algorithms, tied to mathematical optimization, and that can be used on many computational tasks. Machine Learning is a huge technological trend and is now omni present, playing a vital role in diverse fields by using data to train [Shailaja2018] and then to generate knowledge [Reddy2018] by discovering and extracting patterns from subsequently supplied data.

This dynamic is extremely relevant for decision-making on many different practical problem domains and has the potential to provide breakthrough innovation when coped together with the vast amounts of data harnessed by IoT / CPS systems / devices. Machine Learning algorithms can be generally categorized as supervised, unsupervised and reinforcement based. According to [Somvanshi2016], supervised ML algorithms are provided with sample inputs for training data, mapping these inputs to outputs, analyzing, and studying this data, and producing an inferred function that can be used to classify new input data.

Author [Somvanshi2016] also defines unsupervised ML, stating that it is provided with inputs but has no desired output, the classification thus being done with the purpose of correctly differentiating between different supplied datasets. Finally, [Somvanshi16] describes reinforcement ML as actions taken by software to maximize the notion of cumulative reward, being employed in fields such as swarm intelligence and genetic algorithms. The exact same, previously presented

categories are also applicable to the machine learning specialized subfield called known as Deep Learning.

Deep Learning (DL) advances traditional machine learning by addressing some of its known drawbacks. For instance, it reduces the need for the specialized domain knowledge required to feature engineer the massive amounts of data. In other words, DL brings much more automation to the crucial step of input data feature extraction. This in turn allows DL classifiers to perform their task with significantly less human intervention. This is largely possible due to advances in one class of deep learning algorithms, called (Deep) Neural Networks.

Neural Networks (NN) algorithms received such name from their loose inspiration on the way biological brains work and they are called deep because of the multilayer architecture they adopt, consisting of several stacked layers (hence being deep). As pointed out by [Jordan and Mitchell], these algorithms support decision making purposes across many aspects of science, commerce, and government. This is thanks to the capabilities of NNs to generate predictions, in other words, their ability to generate an output correlated to a given input.

The recent growth in the use of deep learning for the enhancement of speech recognition and audio event detection capabilities [Austin2020, Eagle2020, Abdullah2019, Choi2005], especially on safety-driven domains has raised concerns about their robustness against adversarial attacks.

1.1.3 Adversarial Attacks

In the context of machine / deep learning, adversarial attacks happen where the adversary (the attacker) tries to fool the machine / deep learning algorithms being employed under a specific purpose. Some studies have already shown that deep neural network classifiers are susceptible to adversarial attacks aimed at causing misclassifications [Carlini2017, Goodfellow2014].

An adversarial example is defined as a sample of input data which has been modified in a way that is intended to cause a machine learning algorithm to misclassify [Kurakin2016]. Another possibility is to evade detection altogether, in other words, by using adversarial examples, the attacker may fully evade the detection and subsequent correct event classification.

1.2 Related Work and Research Challenges

Research on these detection and classification evasion attacks through adversarial examples have so far largely focused on image-based tasks [Akhtar2018, Athalye2017, Hendrik2017, Su2019]. For the audio processing domain, most of adversarial research focused so far or on speech and speaker recognition applications [Carlini2018, Kwon2019, Abdullah2019, Zhang2017].

On said domain, evasion attacks crafted to fool SR systems usually involve generating malicious audio commands that are recognized and interpreted by the audio models, used in voice processing systems, but that are either inaudible or

that at least have a low degree of perceptibility to the human ear. On the AED front, however, adversarial research has been shy.

Recently, [Subramanian2020] studied the transferability of adversarial attacks in sound event classification, generating adversarial examples based on the Carlini and Wagner attack [Carlini2018]. Some other work has studied countermeasure techniques for improving the resilience of AED systems against adversarial attacks [Roy2018, Carlini2018, Mao2020], however, most of these techniques are passive in nature, besides working on the image space (e.g.: adversarial spectrograms) rather than the audio space.

1.3 Dissertation Contributions

This doctoral dissertation proposes to focus its research efforts on the advance of neural network-enabled audio event detection systems. More specifically, this proposal is about focusing its research efforts on understanding how well state-of-the-art AED real systems work in practice, and later focus on the audio processing portion of AED systems, hence on the study, research and development of audio attack and corresponding defense approaches, applicable to neural networks, tailored for AED tasks.

These networks make up the main component of modern AED systems, and advances on them can directly contribute to a broader spectrum of safer AED applications, thus leading to a better future where these applications, not completely ubiquitous now, but that will, undoubtedly, become ubiquitous in the near future. To reach such advances, this doctoral research was proposed as a

concatenation of three distinct phases, summarized in section 1.4 and shown in Figure 1.

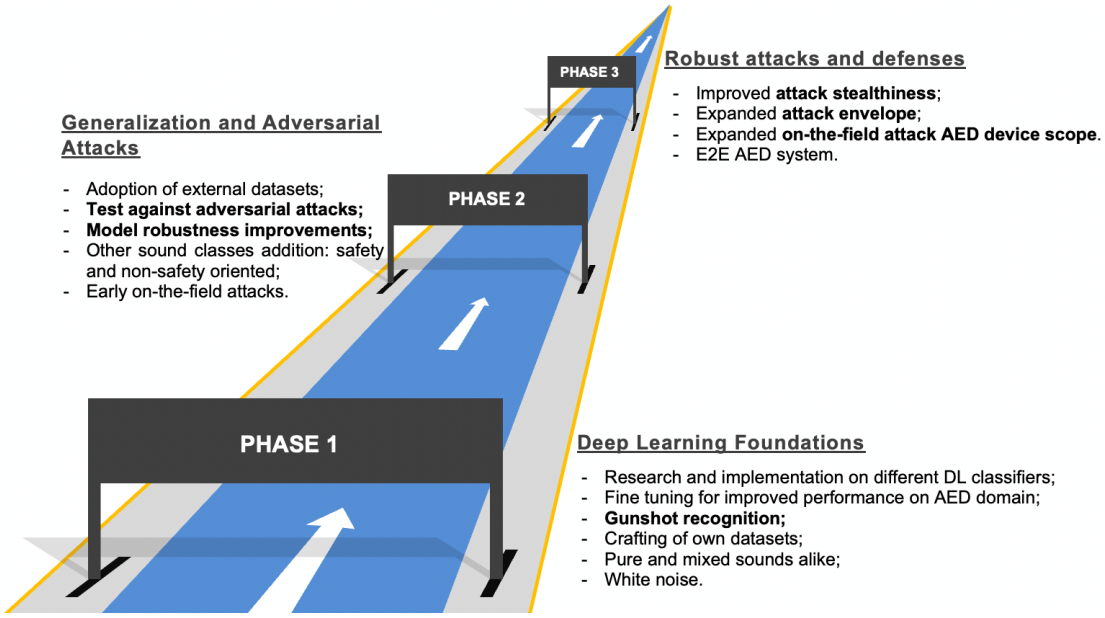

Figure 1: Doctoral research roadmap

1.4 Dissertation Organization

In Chapter 2, the foundation blocks of this doctoral research are presented. Neural network classifiers, tailored for AED purposes, are built, and are attacked by an early form of audio attack, made of white noise. While this early attack is neither inaudible not stealthy, it has several advantages, such as that of being easily reproducible, even by non-technology-savvy individuals, thus being practical for a large roster of adversaries. It also becomes a prime candidate to be employed

as part of concrete attacks, to be carried out in on-the-field fashion against physical AED-capable devices.

In chapter 3, the white noise attacks are once again employed against better refined neural network classifiers, specifically built for AED purposes. These white noise disturbances are now also employed against actual AED-capable gear. More specifically, the disturbances are reproduced, in on-the-field fashion, using audio and computing commodity devices, against AED-capable Google Nest devices, which, as a safety-oriented device, can detect safety-related suspicious sounds, such as those of glass breakage.

Also in chapter 3, existing and actively research defenses are tested out against the white noise attacks. The first line of defense consists of adversarial training, a technique widely employed on the image processing domain, however, used with less intensity in the audio domain. Oversampling, another technique largely employed on the image space is also employed here in the audio space, though to a smaller extent and less relevance than adversarial training.

Also, in chapter 3, an experimental denoising technique, derived from the audio denoising function available in the world-wide known Audacity audio processing tool is tried out as a defense mechanism. An important feature of the experimental implemented function is that unlike in the original, where two audio profiles are needed (one with the audio to be denoised, another one with the known noise to be removed), only a single audio profile is needed and computed for the denoising to take place.

The implementation is said to be experimental because despite bringing at this point relevant improvements to classification results performed on top of denoised audio input, it is known that the denoising algorithm can be further

optimized and should be tweaked in order to generalize better, as currently, it not only brings with it the denoising capability itself, but also brings audio losses, having its effectiveness limited to a few audio classes only. Finally, chapter 3 is closed with different variants of noisy disturbances, as well as signal-to-noise ratio experiments being introduced.

In chapter 4, the attacks against actual AED capable gear are expanded in scope, now including stealthy variants. This is in addition to an extended roster of AED capable black-box devices under test. And E2E AED capable system is also proposed, and this is achieved by coupling previously used in-house built classifiers to external components (i.e.: microphone). As such, evaluations are performed not only against 3rd-party AED devices, but also against a system originated from this research. The E2E AED system outperforms the state-of-the art black box devices while it showcases the shortfalls of adding components to the data pipelines used by DL-enabled AED systems.

CHAPTER 2: DISTUPTING AED CLASSIFIERS WITH WHITE NOISE

In the second chapter of this dissertation, the first phase of this doctoral research and all its early and foundational blocks are introduced. More specifically, in section 2.1 a brief introduction to the research direction is presented, followed by the research motivation and rationale in section 2.2. In section 2.3 the early research methodology is introduced, including detailed information on the implemented state of the art neural networks, suited for, and employed for audio event detection purposes. Details on the planning and design of the early laboratory-level experiments are also provided in this section. In 2.4 the results obtained from the execution of the planned experiments are provided and discussed. A concluding overview of the knowledge generated during the first phase of this research, as well as its next steps are brought in in section 2.5.

2.1 INTRODUCTION

Convolutional Neural Networks (CNN) and Convolutional Recurrent Neural Networks (CRNN), built to be suited for the task of audio event detection and classification, are developed, and tested, initially under ideal conditions. Next, these classifiers are attacked with white noise disturbances, conceived to be simple and straightforward to be implemented and employed, even by non-technology-savvy attackers. The scenario under which these tests and attacks take place is safety-oriented (AED systems tailored to perform safety-related sound detection and classification, such as gunshots).

2.2 Background and Motivation

AED systems for long have been employed for safety purposes, through the detection of suspicious sounds such as gunshots, footsteps, and others. Gunshot sound detection has been extensively researched and represented a good starting point for this doctoral research. AED systems for gunshot detection can be employed anywhere, from home to business, and even public spaces, where they are able to constantly monitor the environment for suspicious events.

These safety-oriented AED systems, just like any other orientation AED system, nowadays make extensive use of state-of-the-art deep learning classifiers, such as convolutional neural networks (CNN) [Zhou2017] and convolutional recurrent neural networks (CRNN) [Lim2017], as their primary detection and classification algorithms. This is due to performance gains these networks generate when compared to legacy approaches.

These gains allowed for widespread CNN and CRNN algorithms employment, reason why this dissertation starts by having both a CNN and a CRNN, tailored for AED purposes, chosen as classifiers to be implemented and employed on the several experiments to follow. The attacks, on this case made of white noise, are employed with the intents of negatively affecting both classifiers' AED performance.

This is because several studies have shown that unwanted noise can have a detrimental effect on neural network classifier performance [Alraddadi2019, Boyat2015]. It is one of the objectives of this doctoral research to study these negative effects brought by such disturbances on deep learning-enabled AED

systems, hence, leading next to the study of countermeasures that can potentially mitigate these negative effects.

2.3 THREAT MODEL

In the adopted threat model (seen in Figure 2), it is assumed that the attacker, while attempting to cause harm, actively adds white noise perturbations to the sound being fed to the AED system. Such threat model was reproduced within the confines of a research laboratory, in other words, white noise was digitally overlaid to the gunshot sounds being used as the AED input. The attacker has no knowledge of the inner implementation details of the AED system.

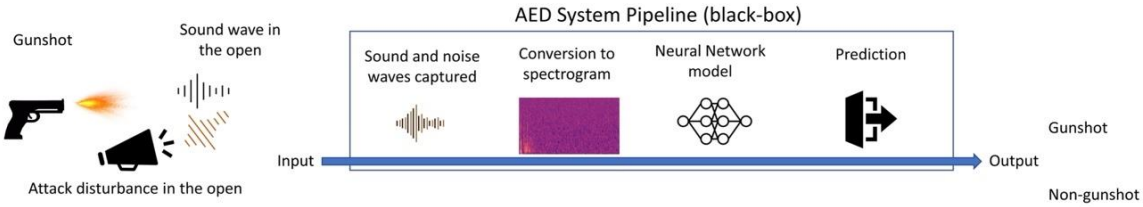

Figure 2: Phase 1 threat model

2.4 Materials and Methods

2.4.1 CNN

Convolutional neural networks are considered to be the best among learning algorithms in understanding image contents [Khan2019]. CNNs were inspired by the organization of the animal visual cortex [Yamashita2018], providing increasingly better performance as they become deeper, while also becoming harder to train [Thakkar2018].

An AED-tailored CNN model based on the work of Zhou et al. [Zhou2017], was implemented. Such tailoring was reached after some initial experimentation, and the final model available for phase 1 of this doctoral research came to be composed by the following components:

- Convolutional layers: three convolutional blocks, each one with two convolutional 2D layers. These layers have 32, 64, 64, 64, 128 and 128 filters (total of 480) of size 3 by 3. Same padding is also applied to the first convolutional layer of each block.
- Pooling layers: three 2 by 2 max pooling layers, each coming right after the second convolutional layer of each convolutional block.
- Dense layers: two dense (also known as fully connected) layers come after the last convolutional block.
- Activation functions: these functions compute the weighted sum of inputs and biases, and as such, are responsible for the firing or no firing

of neurons [Nwankpa2015]. For the presented CNN, ReLU activation is applied after each convolutional layer as well as after the first fully connected layer, while Softmax activation is applied only once, after the second fully connected layer. In other words, ReLU is applied to all inner layers, while Softmax is applied to the most outer layer;

- Regularization: applied in the end of each convolutional block as well as after the first fully connected layer, with 25, 50, 50 and 50% respectively. Regularization, also known as dropout, per [Srivastava2014], addresses the overfitting problem, among other common neural network issues. The CNN uses sparse categorical cross entropy as a loss function and RMSprop as an optimizer. A visual representation of its architecture can be seen in Figure 3.

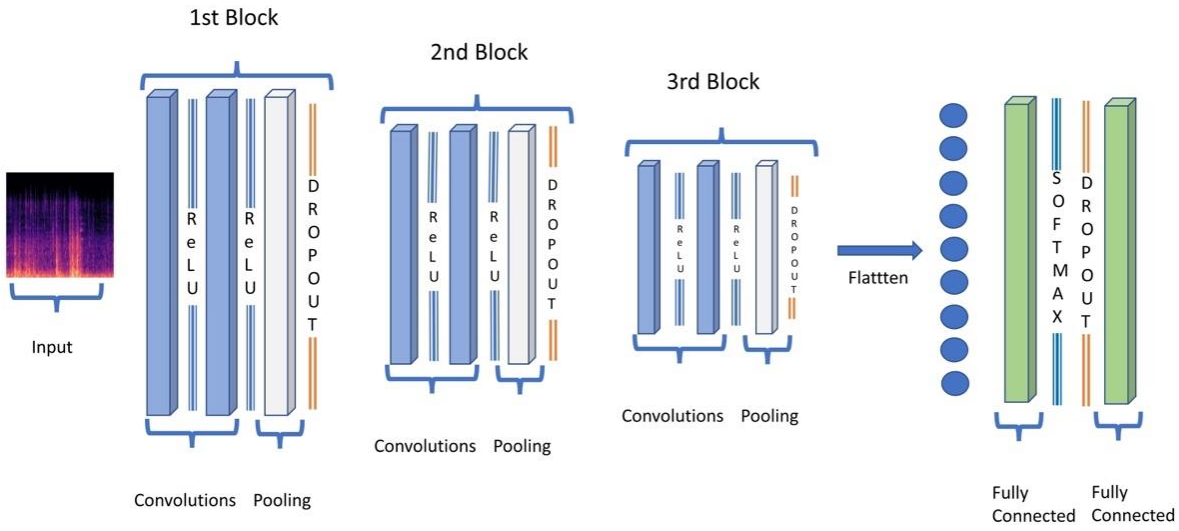

Figure 3: Architecture adopted for Convolutional Neural Network classifier

2.4.2 CRNN

Convolutional recurrent neural networks (CRNN) address one shortfall of regular CNNs - the lack of memory about long-term dependencies over sequences [Fu2017, Gao2018]. Xinyu Fu [Fu2017] proposed to implement a CRNN, where common sigmoid activation function is replaced by a long short-term memory (LSTM) advanced activation. It was shown that CRNNs work better with longer or lengthier inputs [Gao2018] because LSTM activation ensures that outputs of the previous point in time connect to the next point in time.

For this dissertation, a CRNN model was implemented, having been inspired by the work of Lim et al [Lim2017]. Such inspiration comes from the fact that said CRNN has been successfully used for gunshot recognition, presenting reasonable performance. The CRNN model was tailored by considering the results from some initial experimentation, and by the end of the first phase of this doctoral research, came to be composed by the following:

- Convolutional layers: one convolutional block, with one convolutional layer. This block is made by 128 filters of size 32, ReLU activation and batch normalization, pooling layer of size of 40 and a dropout layer of 0.3.
- LSTM layer: one backwards LSTM layer with 128 units, followed by tanh activation and a new dropout of 0.3.
- Dense layers: two stacked dense layers, the first with 128 units and the second with two, each one followed by batch normalization and the first one followed by a ReLU activation and the last one by a Softmax activation.

The CRNN used sparse categorical cross entropy as loss function and Adam as optimizer.

2.4.3 SPECTROGRAMS

The approach adopted by other authors, such as [Zhou2017, Lim2017, Alraddadi2019] with regards to relying to spectrograms as the input of choice to be fed to deep learning models was adopted as part of phase 1 and subsequent phases of this doctoral research. Such common approach, besides facilitating later comparisons, has already proven to be general enough, as spectrograms, being images, fit well as input to both CNN and CRNN models.

Spectrograms display in a graph (usually 2D) the spectrum of frequency changes over time for a sound signal, by chopping it up and then stacking the slices one close to each other [Roelandts2013]. Unlike speech, sound events often have shorter duration but with more distinctive time-frequency representations, which it has been shown to be a good feature for sound classification [Dennis2011, Lim2017, Zhou2017].

One of this doctoral research major design constraints is related to the spectrograms. That constraint resides on the fact that the disturbances must be added to the audio portion of the AED system, and as such, all subsequent spectrogram generation must be free of interference. In other words, the disturbances represented by the proposed white noise attacks are introduced directly to the audio files, prior to their conversion to spectrograms.

As such, in the first phase threat model, the attacker does not have direct access to the spectrogram generation algorithm (black-box IoT/CPS system). This is because it is assumed that the attacker does not have any knowledge about the system and simply tries to alter the sounds generated by the gun before capture by the hypothetical AED system. Samples of spectrograms generated as part of this study can be seen in Figure 4.

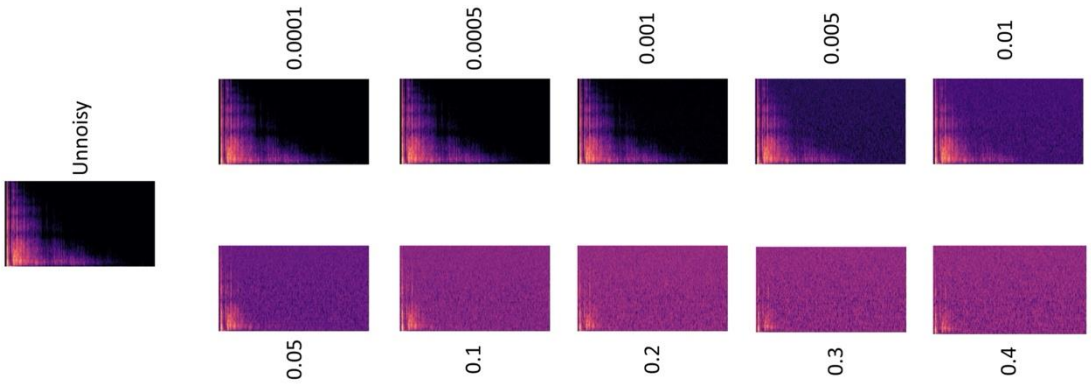

Figure 4: Spectrograms under different noisy conditions

2.4.4 White Noise

As pointed out by [Edmonds2006], white noise happens when “each audible frequency is equally loud”, meaning no sound feature, “shape or form can be distinguished”. A major reason for choosing white noise was the concrete possibility of employing it as part of later practical, on-the-field attacks. Another reason was the simplicity of the attack, thus making it largely available for a large roster of attackers.

In the end, while other types of noise, such as gaussian noise and pink noise generally meet the same criteria previously presented, thus also being prone to be used by an adversary, white noise was the disturbance of choice for phase 1 because this noisy variant is widely adopted by different research across different domains [Dahlan2018, Vashuki2012, Montillet2013].

2.4.5 DATASETS

DCASE 2017 [DCASE2017] provides datasets for several different acoustic classification tasks. For phase 1, the detection of rare sound events dataset was the main dataset acquired and employed in the experiments, since besides being publicly available, it also contains a relatively high number of good quality sounds. To increase the number of data points available to our research, additional gunshot samples from [AirborneSound2017, UrbanSound2017] were also acquired, bringing the total number of samples to 2000.

As positive sounds, 1500 of these sounds are dedicated for training and 500 for testing. Finally, for the negative classes, samples from other sources were acquired, namely [MIMII2019, ESC502021, Freesound2021, Zapsplat2021 and Fesliyan2021]. The negative classes are made of pump, children playing, fan, valve, and music, in other words, sounds that did not carry gunshot sounds. The samples were normalized in terms of frequency, channels, and size (length).

2.4.6 EXPERIMENTS

The experiments conducted as part of phase 1 involve the use of our two neural network classifiers, set as two different representation of an AED system that detects gunshot sounds. Digital gunshot samples are used, first in unnoisy conditions, and then they are infused with progressively higher levels noise. The training and testing sets were crafted within a laboratory, and then employed to train and to test the neural networks, also within a laboratory environment.

The experiments were binary (output could be either *gunshot* or *non-gunshot)*. Both the training and test sets always had the two participating classes in a balanced fashion. In other words, it was guaranteed that each experiment would always have the same number of samples per class in each experiment. A summary of these experiments follows next.

- Unnoisy experiments: Both AED classifiers exposed to digital gunshot sounds, without any disturbance.
- White noise experiments: Both AED classifiers exposed to digital gunshot sounds. The disturbances, made of white noise, are dynamic in nature.

The process for generating the dynamic white noise infused samples can be seen in Algorithm 1. In it, white noise is added to the original audio sample, while again configuring it with the amount of desired noise (through the adjustment

factor or amplitude control), being derived from the highest amplitude already present in the sound sample being disturbed.

Different thresholds for the white noise adjustment factor ranging from 0 (no white noise) and 1 (100 percent noise), thus consisting of multiple thresholds along this interval were tested prior to this beginning of the actual experiments. The initial value (0.0001) was picked based on the criteria of finding a number that was mild in terms of perceptibly while still being able to negatively affect the classifiers.

The choice for the final value (0.4) did not take into consideration the perceptibility criteria, but considered instead a value that would absolutely guarantee, from a practical result perspective, the maximum disturbance to classifier performance possible. The remaining thresholds were chosen in the interval between these initial and final values and the final values were 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4.

```
Result: Perturbed audio sample
initialization;
for number of audio files in the test set do
    sample = load audio file as an array;
    noise = adjustment factor * max element of the array;
    perturbed sample = sample + noise * normal distribution;
    save perturbed sample;
end
```

Algorithm 1: White noise generation.

2.5 RESULTS

When executing the baseline unnoisy experiments, both models perform reasonably well, with accuracies above 80%. This sets the tone for the experiments that come next, where we proceed to attack these same classifiers with white noise. When this happens, both models present drops in classification performance as soon as such noise is introduced to the test sets.

	CNN				**CRNN**			
Condition	**Acc.**	**Prec.**	**Rcl.**	**F1**	**Acc.**	**Prec.**	**Rcl.**	**F1**
Unnoisy	0.88	0.88	0.88	0.88	0.81	0.93	0.66	0.77
0.0001	0.88	0.88	0.88	0.88	0.81	0.93	0.66	0.78
0.0005	0.87	0.89	0.84	0.87	0.81	0.92	0.67	0.78
0.001	0.87	0.89	0.85	0.87	0.81	0.92	0.67	0.78
0.005	0.88	0.90	0.86	0.88	0.81	0.92	0.68	0.79
0.01	0.85	0.90	0.78	0.84	0.81	0.88	0.73	0.80
0.05	0.83	0.90	0.74	0.81	0.84	0.87	0.80	0.83
0.1	0.64	0.93	0.30	0.45	0.70	0.66	0.83	0.73
0.2	0.56	0.94	0.13	0.23	0.66	0.64	0.74	0.68
0.3	0.51	1	0.012	0.02	0.49	0.48	0.35	0.41
0.4	0.5	0	0	0	0.49	0.34	0.11	0.16

Table 1: Phase 1 consolidated results

The drops are small but cumulative, and a sharper drop is noticed when the 0.1 threshold is reached, only to become unacceptably worse from there on, to the point of rendering both models essentially useless. One can also realize that the CRNN is proved to be slightly more robust than the CNN, and one can credit this to its memory advantage over the CNN [Fu2017, Gao2018].

The consolidated results can be seen in Table 1, and the graphical visualization of these results can be seen in Figures 5 and 6.

	0	0.0001	0.0005	0.001	0.005	0.01	0.05	0.1	0.2	0.3	0.4
Acc.	0.81	0.81	0.81	0.81	0.81	0.81	0.84	0.7	0.66	0.49	0.49
Prec.	0.93	0.93	0.92	0.92	0.92	0.88	0.87	0.66	0.64	0.48	0.34
Recl.	0.66	0.66	0.67	0.67	0.68	0.73	0.8	0.83	0.74	0.35	0.11
F1	0.77	0.78	0.78	0.78	0.79	0.8	0.83	0.73	0.68	0.41	0.16

Figure 5: CNN results

	0	0.0001	0.0005	0.001	0.005	0.01	0.05	0.1	0.2	0.3	0.4
Acc.	0.88	0.88	0.87	0.87	0.88	0.85	0.83	0.64	0.56	0.51	0.5
Prec.	0.88	0.88	0.89	0.89	0.9	0.9	0.9	0.93	0.94	1	0
Recl.	0.88	0.88	0.84	0.85	0.86	0.78	0.74	0.3	0.13	0.012	0
F1	0.88	0.88	0.87	0.87	0.88	0.84	0.81	0.45	0.23	0.02	0

Figure 6: CRNN results

2.6 Related Work

This section covers other authors' work related exclusively to phase 1. Several research focusing on audio classification and event detection exist. Some of these relevant works are presented, and the focus mostly stays on "suspicious audio events" detection. These are represented by worrisome sounds such as those of gunshots, glass breaking, crying, etc.

2.6.1 Audio Event Detection

AED systems have been used in environments and applications that have

the capability of collecting real-time audio data multimedia data (including video and/or audio data) and identifying audio events. For example, some health monitoring devices detect sounds, such as coughs to identify symptoms of abnormal health conditions [Matos2006, Larson2011, Peng2009]. Also, some home devices can classify acoustic events of distinct classes (e.g., a baby cry event, music, news, sports, cartoons, and movie) [Matsuoka2020, Vafeiadis2020, Petridis2010, Evangelopoulos2009].

Some commercial initiatives [Shooter2020, Eagle2020, Austin2020, Showen1997] have proposed AED systems specifically designed for gunshot detection. ShotSpotter [Showen1997] and SECURES [Page1995] can detect gunshots by obtaining data from distributed sensors deployed to a large coverage area, and by employing next some traditional signal processing techniques. For instance, SECURES relies on acoustic pulse analysis (pulse peak, width, frequency, shape) performed by electronic circuitry while ShotSpotter employs a specialized software that uses the noise levels in decibels to differentiate gunshots from other sounds.

[Shiekh2017, Busse2019, Rabaoui2008] use Support Vector Machines (SVM) for classifying gunshot sounds. [Shiekh2017] and [Busse2019] differ between each other on the facts that [Shiekh2017] removes reverberations from the audio to make it "cleaner" while [Busse2019] makes use of some oversampling procedures, generating synthetic gunshot sounds. [Chu2004], besides SVM, also uses Gaussian Mixture Models (GMM), a model also employed by [Clavel2005] and [Dufaux2000]. [Chu2004] also goes to the extent of comparing SVM against GMM, the first model outperforming the second for both audio classes used. Finally, [Dufaux2000] besides GMM uses Hidden Markov Models (HMM), which

outperformed GMM on all experiments for all sound classes, under silent and noisy conditions alike.

Other works classify emergency related sounds by employing more modern approaches based on machine learning [Tangkawanit2018, Hansheng2013, Pillos2016, Zhou2017, Khamparia2019], using different variants of Neural Networks (NNs) and different sets of features to perform audio classification. For instance, authors [Zhou2017] and [Khamparia2019] use Convolutional Neural Networks (CNN), while authors [Lim2017] and [Cakir2017] use both CNNs Recurrent Neural Networks (RNN), the former using Long Short Term Memory Unit (LSTM) and the latter using Gated Recurrent Unit (GRU).

An ensemble of CNNs is used by [Donmoon2017] to perform urban sound classification, where two independent, slightly different models take inputs and compute individual predictions, while a final prediction through ensembling both models' probabilities. Author [Ghaffarzadegan2017] uses an ensemble of Deep CNN, Dilatated CNN (DCNN) and Deep RNN for rare events classification.

2.6.2 Images as Part of Audio Event Detection

Most of the neural network-based AED research resort to audio signal transformation into spectrograms, using these images as inputs to their classifiers [Zhou2017, Khamparia2019, Li2016, Donmoon2017]. [Li2016], rather than the spectrograms, uses a combination of partitioned monochrome images derived from spectrograms.

2.7 CONCLUSION

This chapter presented the foundation blocks of this doctoral dissertation, in which research was set to study how deep learning-enabled audio event detection (AED) system work in the first place. Following next, the research was set in a deeper course to study how to make it possible to attack such systems with audio, rather than images, disturbances. White noise was chosen.

The first phase results clearly showed that AED systems, more specifically, the neural networks that power such systems, are susceptible against white noise attacks, as the performance of both the CNN and CRNN classifiers were degraded by nearly 100% when tested against the perturbations. These results are both revealing and important for the following research phases.

Such importance comes from the fact that white noise is simple to reproduce, thus being at the reach of a large roster of potential attackers, including on-the-field capable ones. The same simplicity makes it to also be hard to be filtered out without affecting the sound capture capability needed for an AED system, especially when higher noisy thresholds are used.

Devices embedded with EAD capabilities are already a reality, currently being in the process of becoming ubiquitous for a broad audience. These are real physical devices that employ deep learning models for the detection of suspicious events for security purposes, being largely available for purchase and deployment to homes around the world.

Some examples of these devices are ones manufactured by major companies, such as the Echo Dot and Echo Show by Amazon [Alexa2019], and

Nest Mini and Nest Hub by Google [Nest2021a, Nest2021b]. Despite still being limited in terms of detection capabilities, as most of these devices can detect only a few varieties of audio events, attempts to create general purpose devices, capable of detecting a wide spectrum of audio events, are known to be in the making, e.g., See-Sound [SeeSound2021].

Large scale deployments of these devices as well as attacks against them are as such, just a matter of time. For instance, white-noise reproduction capable gear based on speakers and other specialized equipment are widely available to the broad audience for a long time now [Soundmachines2021]. These can become physical devices that generate audio disturbances on the field.

More sophisticated gear, with increased capabilities are also a reality and are intensively researched and used by military and law enforcement agencies around the world [Mizokami2010, Dormehi2018, Kesslen2019, Chavez2017]. As such, attack solutions that could rely on all sorts of gear, from tiny and discrete devices to large and complex ones are available today.

Therefore, it is not a stretch to envision a scenario where an attacker (e.g.: burglar) could plan for days, weeks or even months in advance on how to deploy attacks against an audio-based AED system. By doing so, such attacker would either delay or avoid detection by an AED system, and as such, gain time to perform his malicious intents.

A burglar could well use an "on-the-field" variant of the white noise attack to disrupt a home-based AED system, thus being able to invade an empty residence without being detected and/or triggering AED-associated alarms and notifications (since a potential glass breakage would not be detected by the under-

attack AED system). After gaining entrance, the burglar could potentially perform his activities, such as robbery, without being disturbed.

Hypothesizing further, as AED systems gain popularity and scale, it is not difficult to envision a scenario where an AED system may be protecting a public area, and terrorists, aware of such monitoring, employ noisy disturbances, like the white noise ones, to disrupt such system. This would make it hard for authorities to respond as an attacker could negate the system's ability to detect a sound of interest (e.g.: gunshots being fired) and subsequently to relay the location of the sound, given such triangulation and notification capabilities are also available.

The hypothesis of a practical white-noise attack that can successfully be applied against exist AED-capable devices, becomes as such, a prime experimentation target for subsequent phases of this doctoral research.

CHAPTER 3: ATTACK AND DEFENSE MECHANISMS

In this third chapter, the second and middle phase of this doctoral research is presented. The chapter starts with its introduction found in section 3.1 and continues to its motivation in section 3.2. The specific methodology followed for this middle part of the research can be seen in section 3.3, including but not limited to the adversarial training strategy as well as to important implementation details regarding the experimental denoising function. Also of key importance are the details brought on the choice of actual AED capable physical devices as well as their arrangement for the first on-the-field research experiments conducted as part of this doctoral dissertation. The experiments aftermath is covered in section 3.4 and the chapter conclusions are covered in section 3.5.

3.1 INTRODUCTION

The second phase of this research maintains its focus on audio event detection (AED) algorithms through the reimplementation of an AED-tailored Convolutional Neural Network (CNN) classifier. Said classifier faces an expanded attack envelope, now consisting of white noise and background noise. Like what happened during phase 1, such development and attacks are conducted in a laboratory level scenario, thus revalidating previously reached results.

Following these lab-level experiments, mainstream commercially available black-box AED-capable devices are set to work as intended, searching for suspicious sounds, namely glass break sounds. These devices, while on duty, are

tested under ideal conditions regarding their AED capabilities and are next exposed to an on-the-field version of the noisy attacks (white and background).

Defenses are also evaluated, and these consist of adversarial training and of an experimental denoising function, applied to the input audio samples, immediately after noise addition and prior to the conversion to spectrograms. In other words, during the second phase of this doctoral research, besides the attacks, defenses are also applied and studied.

3.2 Background and Motivation

Phase 1 of this doctoral research was important because its results clearly showed that it is possible to use simple, yet effective white noise to disturb deep learning (DL) classifiers employed for audio event detection (AED) tasks. This early research, however, brings with it important limitations, such as the use of a single type of disturbance, of a single audio class of interest (gunshot), and the fact that it was entirely confined to a laboratory.

Another very important limitation was the lack of experimentation regarding possible countermeasures to the white noise attacks, even if held in a laboratory scenario. Phase 2, as such, addresses these limitations, since it was devised to continue testing the robustness of multiple security critical AED tasks, implemented as CNNs classifiers, but also to test existing third-party Nest devices, manufactured by Google, which run their own black-box deep learning models.

The adversarial examples (audio perturbations) were made of white and different forms of background noise. Despite the increased attack envelope, the

disturbances in the scope of phase 2 remain easy to create, and to reproduce, being at the grasp of many potential attackers, hence, this important research design constraint conceived back during phase 1 remains valid.

In addition to the attacks, the improvement of classifier's robustness through specific countermeasures are also studied. These consisted of both adversarial training and audio denoising, and they are evaluated, both in isolation from each other as well as in combined fashion, while being applied to the audio input fed to both the in-house built CNN classifier and to the third-party device.

Given the several critical applications of AED systems and vast collection of possible usage scenarios for these AED systems, during phase 2 of this research the possible scenarios are narrowed down, thus a home security scenario was selected to emulate a physical world AED deployment, where the AED system would be constantly monitoring the environment for suspicious events.

Considering that the Nest devices come from the factory being capable of detecting glass break sounds, said sound class plus the original gunshot class from phase 1 were chosen to be phase 2 positive classes (the ones containing the sounds of interest. In the revised threat model for phase 2, the AED system is deployed as part of a home security system, and the adversary, while attempting to cause harm, aims to prevent the AED system from correctly detecting and classifying the sound events. For this purpose, the adversary generates some noise (e.g., background noise or white noise) which can perturb the audio being captured by the AED system.

3.3 Threat Model

In the evolved threat model for phase 2, the attacker employs not only white noise, but also background noise to the sounds used as input to the AED system. Besides additional lab-level experiments, now the attacker also performs the attacks while being on the field, using commodity equipment to perform the attack. The attacker remains bling to the inner implementation details of the AED system, regardless of if it made of in-house built AED classifiers or third-party AED capable devices.

3.4 Materials and Methods

3.4.1 CNN AED Classifiers

New tests of in-house built (for AED purpose) CNN classifiers are conducted to validate and to confirm Phase 1 results. While modern AED systems are made up of a pipeline containing multiple components, it is straightforward to reason that several of these systems have in their embedded neural network models, one of the most, if not the most critical component enabling the delivered AED capabilities. This makes these classifiers into a prime component for testing.

Regardless of which results are to be achieved when testing end-to-end AED systems (as explained in section 3.4.2), it is likely that whatever result is to be achieved, it will be the result of cumulative (good or bad) performance by each

component making up the entire system (quality of the embedded microphones, capabilities of the local machine learning chips, compromises needed for the on-device deep learning models etc.).

Given this research focus on software rather than hardware, it is only natural that the focus on singling out the neural network individual component for testing. As such, Convolutional Neural Networks, capable of detecting gunshot and glass break sounds are once again tested under ideal conditions (noise-free) and "under-attack" conditions, when they are fed with digitally disturbed audio samples, thus emulating what would be found during actual on-the-field audio attacks.

3.4.2 THIRD-PARTY AED DEVICES

In addition to the CNN classifier in a laboratory environment, a second testing arena was added to the roster of experiments being carried out as part of phase 2, now including third-party AED capable devices. The devices selected were the ones which were readily available for purchase in the open market, in other words, were largely at reach of both customers as well as potential attackers. The overall research framework can be seen in Figure 7.

Given the well-known involvement of Google with Deep Learning (e.g., creation and release of TensorFlow), and the fact that Google AI-enabled devices, including Nest devices are already widely used in day-to-day life [Policyadvice2021], the following devices were selected:

- Nest Mini: from the large variety of Nest devices available, he most basic device possible were chosen, namely the Nest mini

[Nest2021a]. The Nest mini device, currently in its second generation [Analyticsindiamag2019], and already includes a machine learning chip capable of implementing advanced techniques such as natural language processing and speech recognition. Yet another advantage of these devices is the fact that they can work in pairs, in theory augmenting their detection capabilities.

- Nest Hub: this is defined in [Nest2021b] device, and offers all Nest mini capabilities, besides a display [Pocketlint2021]. Nest hub can be an attractive device to consumers who want to start their own smart home implementation with some simplicity but want something more refined and capable than the simple Nest mini.

3.4.3 Adversarial Attacks and the Addition of Background noise

Two variants of evasion attacks based on noise were selected and implemented for this phase, namely white and background noises. Both can be used to generate fast and straightforward perturbations, as in a lab scenario, with the aid of commodity computers, up to 2,000 digital noisy samples can be generated per minute. They can also be easily employed as part of on-the-field attacks, using, for example, commodity sound speakers.

With a different profile from the flat, absent of features profile provided by the already previously successfully used white noise, background noise, on the other hand, is represented by all sorts of common noise occurring during the normal course of day-to-day business, and that may overlay to any sound of

interest. Examples of such noise would be that of people talking, active vehicle traffic, music playing, etc.

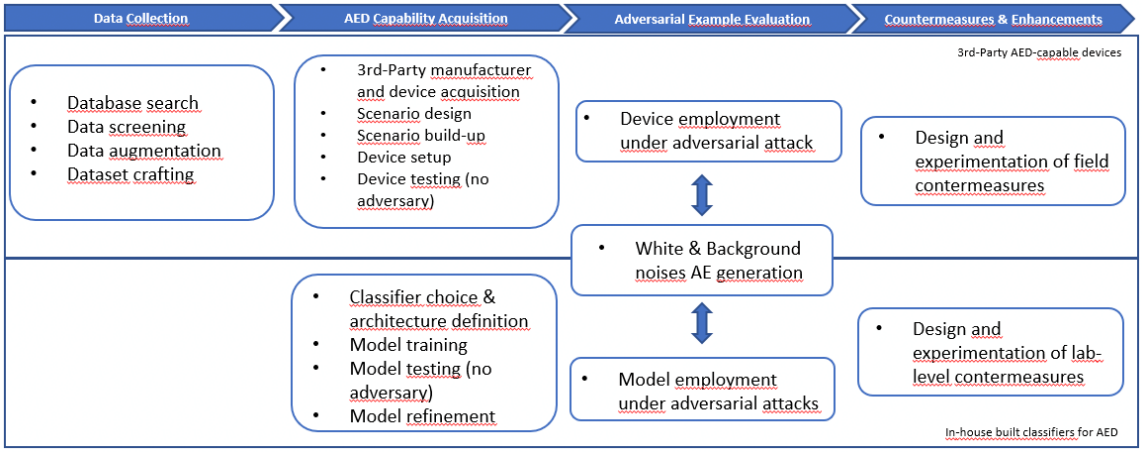

Figure 7: Research execution framework

The same way that white noise was added to the audio samples prior to their conversion to spectrograms, the same holds true for background noise. Also, important to mention is that this is true for both the in-house built classifier testing as well as for third party AED capable gear testing, in other words, the disturbances are added when these devices are actively listening for glass break sounds. On the case of the latter, this is done through loudspeakers.

Algorithm 2 shows the mechanism for the addition of background noises to a given audio sample. In it, two separate files are retrieved, one with the sound of interest, and one with the background noise. Such background noise is added to the sound of interest without any modification other than the one introduced by adjustment factor, which simply controls the amplitude (or loudness) of the noise.

```
Result: Perturbed audio sample
initialization;
for number of audio files in the test set do
    sample = load audio file as an array;
    noise = load audio file as an array;
    adjusted noise = noise + adjustment factor;
    perturbed sample = sample + adjusted noise;
    save perturbed sample;
end
```

Algorithm 2: Background noise generation.

Besides adversarial attacks against AED systems, phase 2 of this doctoral research adds investigative efforts on techniques for increasing the robustness of these systems against adversarial examples. Three techniques were implemented and later evaluated through additional defensive experiments: oversampling, adversarial training, and audio denoising. The rationale behind these techniques is presented next.

3.4.4 OVERSAMPLING

Overfitting is known to be related to adversarial sensitivity and some works have shown that mitigating overfitting improves the accuracy on adversarial examples [Kubo2019, Galloway2018]. CNNs classifiers are known to be prone to overfitting, when “deep” (having multiple layers) architectures are used, and when a class imbalance exists (a class having more samples than the others), affecting convergence during the training phase and generalization of a model on the test set [Buda2018].

Oversampling is one of the most popular augmentation techniques [Shijie2017, Perez2017] that can mitigate overfitting [Buda2018]. One of its forms

consists of applying pure sample duplication, without modifications to the duplicates [Wei2005]. This is the approach adopted, in other words, oversampling by cloning, thus increasing the number of data samples in the training sets.

It is important to highlight that the reports of the oversampling results are omitted from this dissertation report due to too small gains generated by this technique. Such small improvements come not from a lack of effectiveness by oversampling per se, but from the fact that at phase 2, the classifiers have been improved to a point that their performance is good to the point of making it hard to have them benefiting from oversampling in a significant way.

3.4.5 Adversarial Training

This is a popular technique applied by several researchers [Wang2019, Song2018]. It consists of introducing some adversarial examples into the training set, thus leading to increased resilience against adversarial attacks through learning directly from adversarial examples. While adversarial learning has been mostly been used for image classification tasks, this research applies it to audio.

3.4.6 Audio Denoising

Audio denoising techniques exist to remove or at least to mitigate the noise existing in an audio sample. Several works have used filters to perform audio denoising, thus leading to improvement in classifier's performance. Some works

[Kiapuchinski2012, Hodgson2010, Audacity2020] used some variation of a technique called Spectral Noise Gating [Hodgson2010].

Such work consists of performing the reduction of a signal found to be below a given threshold (the noise), and an important point about it was brought up by [Kiapuchinski2012], consisting of its requirement to have a noise profile (extracted from the known noise), from which a smoothing factor will be derived and applied to the signal that requires denoising (the whole sound).

Following a similar approach, as part of this dissertation, a custom experimental denoising spectral gating function was implemented, being based on the noise reduction function employed by [Audacity2021], the open-source digital audio editor and recording application software and rewritten in python code by [Sainburg2018]. Despite the similarities to the original base version, the in-house built version of the denoising function bears important modifications to the original.

For instance, while in the original function, as explained before, two input sounds are required for the denoising to take place (one with noise, one with the audio to be denoised), the experimental function uses a same “whole sound” as both audio and noise profiles donor, hence not requiring two separate inputs. This is because in this research threat model, the defender does not have any knowledge about the noise function used by the adversary.

Besides the previously mentioned difference with regards to audio inputs, the experimental implementation also brings additional changes, for instance, ones related to frequency channels}, Fourier transform frames, window, and hop lengths, and time and frequency smoothing filter setups. The pseudo-code for the denoising function can be seen in algorithm 3.

```
Result: Perturbed audio sample
initialization;
for number of audio files in the test set do
    sample = load audio file as an array;
    noise = load audio file as an array;
    sample profile = calculate statistics specific to sample;
    noise profile = calculate statistics specific to noise;
    if sample profile < noise profile then
        apply smoothing filters;
        save denoised sample;
end
```

Algorithm 3: Denoising mechanism.

These changes, even though they are not final (as they could be further improved in the next iterations of this research) are fundamentally in the right direction. This is because the modified algorithm can reduce the noise fingerprint on each frequency spectrum of the audio, while at the same time representing a better tailored approach for the AED domain problem at hand.

3.4.7 DATASETS

Besides the audio sample sources of phase 1, several other public audio databases were chosen to be the source of the audio samples used on the several experiments from phase 2. These databases were:

- Detection and Classification of Acoustic Scenes and Events or DCASE dataset [DCASE2017]: From 2017 and 2018 editions, the DCASE datasets include normalized audio samples with a single instance of an event of interest happening anywhere inside each

audio sample of 30 seconds in length, hence the “rare" denomination. Each sample is created artificially and has background noise made of everyday audio.

- Urban Sounds Dataset [UrbanSound2017]: A database made of everyday sounds found at urban locations. The samples are not normalized and vary quite a bit among themselves.
- MIMII Dataset [MIMII2019]: A dataset conceived to aid the investigation and inspection of malfunctioning industrial machines.
- Airborne Sound [AirborneSound2017]: An open and free database with audio samples destined to be employed on different sound effects. One such case is that of guns and medieval weapons. The gun part has high quality audio on several different types of guns, recorded from different positions.
- Environmental Sounds [ESC502021]: A dataset of 50 different sound events and over 2,000 samples.
- Zapsplat [Zapsplat2021]: Over 85,000 professional-grade audio samples as royalties-free music and sound effects.
- FreeSound [Freesound2021]: A collaborative database of Creative Commons Licensed sounds.
- Fesliyan Studios [Fesliyan2021]: A database of royalty-free sounds.

The samples from these datasets that contain audio events of interest (security/ safety related) are called as “positive samples”, and those that do not contain sounds of interest as “negative samples”. These samples, before being used in the experiments, were cleaned, and preprocessed in the following ways:

- Frequency Normalization: where the frequencies of all samples are normalized to 22,000 Hertz, to be within the human audible frequency.
- Audio Channel Normalization: where needed, the number of channels of all samples were converted from stereo to monaural, as it is easier to find new samples bearing a single channel.
- Audio Length Normalization: where all samples with less than 3 seconds in length were discarded.

3.4.8 EXPERIMENTS

The several experiments of phase 2 are listed next:

- Experiment 1: the baseline experiments (the ones executed under unnoisy, ideal conditions).

    - Experiment 1a - Binary CNN Classifiers: four binary models are trained, each with 1000 positive samples and 1000 negative samples. The positive samples in each model belong to one of the categories of sounds, i.e., dog bark, glass break, gun, and siren. The negative portion of the training set was kept unaltered throughout the 4 experiments and was made of a combination of 200 samples of each one of the five different negative classes previously presented.

The respective test sets were made of 300 samples, 150 positives, and 150 negatives.

- Experiment 1b - Multiclass CNN Classifier: this experiment involved a multiclass version of the CNN algorithm, including now all 4 positive classes at once. Its goal is to investigate if multiclass classifiers provide different results or show different behavior compared to binary classifiers, even though currently readily available AED systems are dedicated to detecting one or two audio classes only. In this experiment, the training sets were made of the 4,000 positive samples used in Exp 1a, with no negative classes.

- Experiment 2: the testing of third-party AED-capable devices, as seen in Figure 8. Also included in this batch is the testing of a multiclass version of the binary classifiers used extensively up to this point, as there was the need to confirm if the base classifier architecture would work well under binary and multiclass conditions alike.

    - Experiment 2a - Digital Pure Audio Inputs: 3rd-party devices exposed to digital glass break sounds, without any disturbance being played through the loudspeakers.
    - Experiment 2b - Real Pure Audio Inputs: 3rd-party devices exposed to real glass break sounds, without any disturbance being played through the loudspeakers.

- Experiment 2c - Background Noise Disturbed Inputs: 3rd-party devices exposed to real glass break sounds, with background noise disturbance being played through a loudspeaker.
- Experiment 2d - White Noise Disturbed Inputs: 3rd-party devices exposed to real glass break sounds, now with white noise disturbance being played through a loudspeaker.
- Experiment 2e - Binary CNN Classifier and Pure Glass Break Recordings: The CNN classifiers, now being fed, during test phase, with glass break sounds recorded during experiments 2a, 2b and 2c, by the S10+ and S20 Ultra devices.

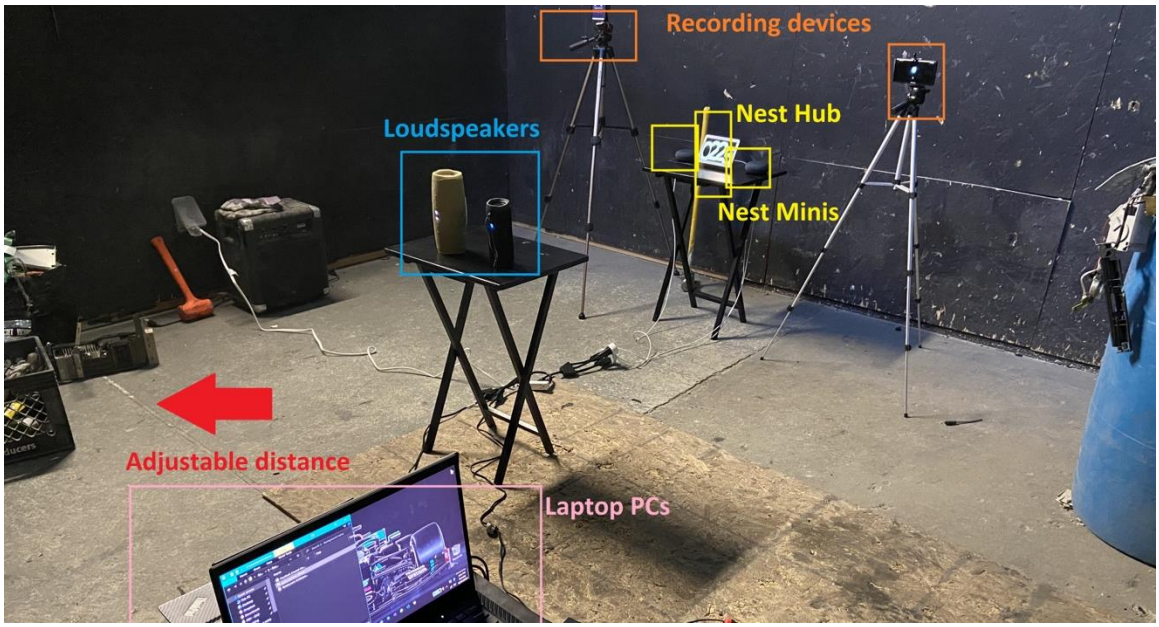

Figure 8: Early on the field experiments

Experiment 3: Adversarial Examples: Test of the same two respective, previously trained gunshot and glass break classifiers, against increasing levels of

background and white noises. For the background noise, Pydub python library was used to digitally add two different background noises, namely car traffic and people talking, to the test set samples to be fed to the models.

To clarify, these background noises are not related to the negative classes that used to train and test the classifiers. Therefore, if the models misclassify the adversarial samples generated via background noise, it is not due to existence of similar samples in the negative class. The positive classes are made of gunshot and glass break.

The signal-to-noise ratio was kept at 10 decibels (measured on site), similarly to the on-the-field experiments on third-party devices. The Numpy python library was used to digitally generate white noise disturbances, as well as Librosa and SoundFile libraries to add the disturbances to the test set samples. Eleven different test sets, each having 100% of their samples overlaid with 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4 and 0.5 white noise levels were created.

- Experiment 3a - Glass break Classifier and Background Noise Infused Audio Inputs: Glass break classifier from experiment 1, tested against three different test sets, having 25%, 50% and 100% of their samples infused with background noise.
- Experiment 3b - Gunshot Classifier and Background Noise Infused Audio Inputs: Gunshot classifier from experiment 1, tested against three different test sets, having 25%, 50% and 100% of their samples infused with background noise.
- Experiment 3c - Glass break Classifier and White Noise Infused Audio Inputs: Glass break classifier from experiment 1, tested against the eleven different white noise infused test sets.

- Experiment 3d - Gunshot Classifier and White Noise Infused Audio Inputs: Gunshot classifier from experiment 1, tested against the eleven different white noise infused test sets.

Experiment 4: Background Noise for Training

The objective is to test effectiveness of adversarial training as a countermeasure against evasion attacks, when background noise-infused samples are added to training sets.

Experiment 4a: Glass Break with Background Noise:

- From Experiment 3a, its 100 percent background noise infused glass break test set is reused, however its train set is modified, now turning 100 percent of its samples into adversarial examples by infusing them with background noise.
- Experiment 4b: Gunshot with Background Noise: from Experiment 3b, its 100 percent background noise infused gunshot test set is reused, however its train set is modified, now turning 25, 50 and 100 percent of its samples into adversarial examples by infusing them with background noise.

Experiment 5: White Noise Adversarial Training:

The objective is to test the effectiveness of adversarial training as a countermeasure to evasion attacks. In other words, white noise infused samples are now added to the train sets.

- Experiment 5a - Glass break with White Noise: the eleven glass break test sets from Experiment 3c are reused, while the glass break train sets from Experiment 1a are modified, having added to it, proportionally, ten out of the eleven white noise levels previously used (0.0005 to 0.5). As such, every white noise level had 100 samples included in 6a train set.
- Experiment 5b - Gunshot with White Noise: the eleven gunshot test sets from Experiment 3d are reused, while the gunshot train set from Experiment 1a are modified, having added to it, proportionally, ten out of the eleven white noise levels previously used (0.0005 to 0.5). As such, every white noise level had 100 samples included in 6b train set.

Experiment 6: Denoising Background Noise:

The objective is to test the effectiveness of the experimental denoising in the face of background noise attacks.

- Experiment 6a - Glass break Test sets: the original free-of-noise glass break train set from experiment 1a is reused, while the 100%

background noise infused test set from experiment 3a is denoised and reused.

- Experiment 6b - Gunshot Test sets: the original free-of-noise gunshot train set from experiment 1a is reused, while the 100% background noise infused test set and from experiment 3b is denoised and reused.

Experiment 7: Denoising White Noise

The objective is to test the effectiveness of the experimental denoising in the face of white noise attacks.

- Experiment 7a - Glass break Test sets: the original free-of-noise glass break train set from experiment 1a is reused, while the eleven glass break white noise infused test sets from experiment 3c are denoised and reused.
- Experiment 7b - Gunshot Test sets: the original, free-of-noise gunshot train set from experiment 1a is reused, while the eleven-gunshot white noise infused test sets from experiment 3d are denoised and reused.

Experiment 8: SNR and Other Types of Noise

This additional test was performed after all other experiments have been completed, under the intentions of a) to verify the possibility of using other types of noise, beyond white and background, due to possible differences in natural stealthiness; and b) to verify how to use Signal-To-Noise-Ratio (SNR), as the main

measure of power difference between the audio to be disturbed and the disturbances themselves. Employing SNR means that future research results could be reported using an industry standard, thus replacing the noisy thresholds used thus far.

Since a new implementation followed by some tests would be needed to assess how effective the SNR-based disturbances would be, this was also a good opportunity to assess new types of noise, beyond the white and background ones. This is because every practical application that deals with audio signals also deals with the issue of noise.

As stated by [Prasadh2017], “natural audio signals are never available in pure, noiseless form”.' As such, even under ideal conditions, natural noise may be present in the audio being in use. Just to cite a few, some common types of noise are:

a) Gaussian noise: arising n amplifiers or detectors, having a probability density function that is proximal to real world scenarios [Rajaratnam2018].
b) Gaussian noise: distributed in a normal, bell-shaped like fashion.
c) Pink noise: also known as flicker noise, is a random process with an average power spectral density inversely proportional to the frequency of the input signal [Isar2016].
d) Cauchy noise: similar to gaussian noise and its bell-shaped curve, the Cauchy noise distinguishes itself by presenting a density function with a shape that has a higher density at center and also has a longer tail [Ito2016].

While all these noises can be used by an adversary, the types of noise chosen for these final experiments were pink, brown, and blue, as these variants are largely available for download as standalone audio samples. This is important, as the SNR function created (using librosa and numpy libraries) for these experiments was put together in a way to make it as straightforward as possible to generate the SNR-based disturbances.

This has been achieved by providing the function with a sample to be disturbed, a disturbance sample, and an SNR threshold to be achieved. The energies of both samples are calculated, and the increase or decrease needed to reach the specified SNR threshold is applied to the noise signal directly. By using standalone disturbance samples, this also means that the original noises (white and background) are compatible with this new function and can be reused in future SNR-based experiments.

3.5 RESULTS

Table 2 shows that the base classifiers, trained only on noise-free samples, present great performance. The four binary classifiers, namely dog barking, glass breaking, gunshots and siren, all perform above 94% accuracy, while the multiclass classifier that includes all these same classes at once, also performs well, having an accuracy of close to 93%. Therefore, the multiclass classifier is on par with the binary classifiers.

AED System	Experiment Id.	Train Samples	Test Samples	Acc.	Prec.	Rcl.	F1
Custom CNN	1a – Bark – Digital	2000	300	0.96	0.96	0.96	0.96
	1a - Glass break – Digital	2000	300	0.99	0.99	0.99	0.99
	1a – Gun – Digital	2000	300	0.99	0.99	0.99	0.99
	1a – Siren – Digital	2000	300	0.94	0.94	0.94	0.94
	1b – Multiclass - Digital	4000	600	0.93	0.93	0.93	0.93
	2e – Glass break - Real	2000	150	1	1	1	1

Table 2: CNN baseline results

As it can be seen from Experiments 2a to 2d in Table 3, even under unnoisy conditions, the Nest devices perform poorly, with a detection rate of about 33%, which only gets worse when disturbances are introduced to the environment. Particularly, the background noise can reduce detection rates by 22% while white noise reduces them by 25%. This is concerning as families may trust their security and safety to these devices to some extent.

Absent from the table is information about the configurations of devices used (isolated or in combination under separate distances), as experiments were conducted under different distance setups, however, even though these

differences were considered during experiment design, so much so that different setup experiments were conducted, it was not possible to obtain any significant distinct performance differences from these setups.

AED System	Experiment Id.	Attempts	Detected	Missed	Detection Success Rate
3rd-Party Nest Devices	2a – Glass break – digital	15	11	15	11%
	2b – Glass break (unnoisy) – digital	18	6	12	33%
	2c – Glass break & BN - real	18	2	16	11%
	2d – Glass break & BN - real	12	1	11	8.3%

Table 3: Tests with 3rd-Party AED capable devices

Finally, as part of experiment 2e, a subset of the real glass break sounds recorded by the S10 and S20 devices (75 in total) were used to test the previously in-house trained glass break CNN classifier. Under these circumstances, the CNN model had an even higher detection accuracy, now of one hundred percent.

Experiments 3a and 3b are based on background noise as an attacking mechanism. As such, from Experiment 1a, the glass break and gunshot baseline classifiers as well the test sets are reused, except that these sets were modified by progressively increasing the number of samples within them that are infused with background noise. The results of these experiments can be seen in Table 4, which shows the effectiveness of the background noise disturbances, as they increasingly affect classifier's performance.

The results produced are not even, since the glass break classifier performs worse to the disturbances, presenting an accuracy drop of up to 28% when 100% of the test set is infused with background noise. Note that the noise is added to only the samples in the positive class, e.g., gun, glass break. In contrast, the gunshot classifier has its performance dropping by around 7%.

Different performance drops for different classes due to background noise infusion was expected, as the effectiveness of these disturbances will be affected by several factors, for instance, how feature rich the sound of interest is to begin with. For now, this is pointed as the primary reason for the difference on these experiments involving gunshot and glass break (the first being much louder and distinct than the second).

The same approach is adopted during previous Experiments 3a and 3b, and as such the glass break and gunshot baseline classifiers as well their test sets are reused, but now all test samples are infused with progressively higher white noise levels, ranging from 0.0001 to 0.5. The whole list of white noise levels as well as the experiment results are disclosed in Table 4.

Based on these results, the gunshot sounds prove to be more susceptible to the white noise disturbances than glass break, presenting sharp accuracy drops

of over 40%. Still, glass break does not lag much behind, showing drops close to 40%. It can be observed that white noise-infused adversarial examples significantly decrease the performance of both the gunshot and glass break classifiers, but not that of glass breaking classifier.

The effectiveness of the selected countermeasures against evasion attacks have been tested next. The defensive techniques employed rely on adversarial training, where some adversarial examples are added to the training sets. Experiments 4a and 4b examine adversarial training using samples with background noise. The baseline glass break and gunshot training sets from experiment 1a are reused, however, they are modified by being infused with background noise, having 100% of their positive samples modified in this way.

Two extra experiments were also created, combining the original free of noise train sets to a fully disturbed train set. Similarly, experiments 4c and 4d take and modify the baseline 1a train sets, however ten out of eleven white noise levels (from 0.0005 to 0.5) are added proportionally to the train sets, each level, thus, perturbing two hundred samples.

The retrained models are tested against the same eleven white noise infused test sets seen at experiments 3c and 3d. Table 5 shows the results for these experiments. For adversarial training using sample with background noise, up to 8% and 29% improvement for gunshot and glass break are achieved, respectively, which is a significant result.

AED System	Experiment	Acc.	Prec.	Rcl.	F1
Glass Break	Baseline (1a)	0.99	0.99	0.99	0.99
Gunshot	Baseline (1a)	0.99	0.99	0.99	0.99
Glass break - digital	3a – 25% BN	0.88	0.90	0.88	0.87
	3a – 50% BN	0.76	0.84	0.76	0.75
	3a – 100% BN	0.71	0.82	0.71	0.69
Gunshot - digital	3b – 25% BN	0.96	0.93	0.96	0.96
	3b – 50% BN	0.94	0.95	0.94	0.94
	3b - 100% BN	0.92	0.93	0.92	0.92
Glass break - digital	3c – 0.0001 WN	0.99	0.99	0.99	0.98
	3c – 0.0005 WN	0.95	0.96	0.95	0.95
	3c – 0.001 WN	0.95	0.95	0.95	0.94
	3c – 0.005 WN	0.98	0.98	0.98	0.98
	3c – 0.01 WN	0.99	0.99	0.99	0.98
	3c – 0.05 WN	0.93	0.93	0.93	0.93
	3c – 0.1 WN	0.81	0.86	0.81	0.80
	3c – 0.2 WN	0.81	0.86	0.81	0.80
	3c – 0.3 WN	0.66	0.8	0.66	0.62
	3c – 0.4 WN	0.65	0.79	0.65	0.60
	3c – 0.5 WN	0.61	0.78	0.61	0.54
Gunshot - digital	3d – 0.0001 WN	0.98	0.98	0.98	0.98
	3d – 0.0005 WN	0.85	0.88	0.85	0.84
	3d – 0.001 WN	0.9	0.92	0.9	0.9
	3d – 0.005 WN	0.66	0.8	0.66	0.66

	3d – 0.01 WN	**0.63**	**0.79**	**0.79**	**0.57**
	3d – 0.05 WN	**0.59**	**0.77**	**0.59**	**0.5**
	3d – 0.1 WN	**0.58**	**0.77**	**0.58**	**0.49**
	3d – 0.2 WN	**0.55**	**0.76**	**0.55**	**0.43**
	3d – 0.3 WN	**0.54**	**0.76**	**0.54**	**0.41**
	3d – 0.4 WN	**0.52**	**0.76**	**0.52**	**0.38**
	3d – 0.5 WN	**0.5**	**0.75**	**0.5**	**0.34**

Table 4: CNN adversarial attack tests

For adversarial training based using samples with white noise, nearly 50% improvement for both gunshot and glass break are achieved. The retrained models are tested against test sets explained in experiments 3a and 3b, where 100% of their positive samples disturbed by background noise. Finally, as the final defense mechanism, denoising the adversarial test sets through the experimental denoising function is attempted.

Experiments 6a and 6b involve denoising the 100% background noise infused test sets from experiments 3a and 3b, while experiment 7a and 7b involve denoising the ten white noise infused test sets from experiments 3c and 3d. The train sets are the baseline ones from Experiment 1a. The denoising consolidated results can be seen in Table 6.

AED System	Experiment	Acc.	Prec.	Rcl.	F1
Glass break	4a – 100% BN	1	1	1	1
Gunshot	4b – 100% BN	1	1	1	1
Glass break	5a – 0.0001 WN	0.99	0.99	0.99	0.99
	5a – 0.0005 WN	0.99	0.99	0.99	0.99
	5a – 0.001 WN	0.99	0.99	0.99	0.99
	5a – 0.005 WN	1	1	1	1
	5a – 0.01 WN	1	1	1	1
	5a – 0.05 WN	1	1	1	1
	5a – 0.1 WN	1	1	1	1
	5a – 0.2 WN	1	1	1	1
	5a – 0.3 WN	1	1	1	1
	5a – 0.4 WN	1	1	1	1
	5a – 0.5 WN	1	1	1	1
	5b – 0.0001 WN	0.98	0.98	0.98	0.98
	5b – 0.0005 WN	0.98	0.98	0.98	0.98
	5b – 0.001 WN	0.99	0.99	0.99	0.99
	5b – 0.005 WN	0.99	0.99	0.99	0.99
	5b – 0.01 WN	0.99	0.99	0.99	0.99
	5b – 0.05 WN	0.997	0.997	0.997	0.997
	5b – 0.1 WN	0.997	0.997	0.997	0.997
	5b – 0.2 WN	0.997	0.997	0.997	0.997

	5b – 0.3 WN	0.997	0.997	0.997	0.997
	5b – 0.4 WN	0.997	0.997	0.997	0.997
	5b – 0.5 WN	0.997	0.997	0.997	0.997

Table 5: CNN adversarial training defensive tests

Experiment 7a achieves nearly 3% accuracy improvement for both background noise denoised gunshot and glass break, while 7b achieves over 7% improvement for white noise denoised gunshot. Experiment 10a also achieves up to a low 1% improvement for glass break. Despite the modest improvements seen during the execution of the denoising experiments and its corresponding results, the denoising experiments showcase the benefits of the proposed spectral gating denoising technique, especially developed in the scope of this doctoral research.

AED System	**Experiment**	**Acc.**	**Prec.**	**Rcl.**	**F1**
Glass break	6a – 100% BN	0.74	0.83	0.74	0.72
Gunshot	6b – 100% BN	0.94	0.95	0.94	0.94
Glass break	7a – 0.0001 WN	0.99	0.99	0.99	0.99
	7a – 0.0005 WN	0.97	0.97	0.98	0.98
	7a – 0.001 WN	0.95	0.96	0.95	0.95
	7a – 0.005 WN	0.97	0.97	0.97	0.97
	7a – 0.01 WN	0.97	0.97	0.97	0.97
	7a – 0.05 WN	0.57	0.7	0.57	0.49

	7a – 0.1 WN	0.96	0.98	0.96	0.96
	7a – 0.2 WN	0.98	0.98	0.98	0.98
	7a – 0.3 WN	0.98	0.98	0.98	0.98
	7a – 0.4 WN	0.98	0.98	0.98	0.98
	7a – 0.5 WN	0.98	0.98	0.98	0.98
Gunshot	7b – 0.0001 WN	0.98	0.98	0.98	0.98
	7b – 0.0005 WN	0.85	0.88	0.85	0.84
	7b – 0.001 WN	0.91	0.92	0.91	0.91
	7b – 0.005 WN	0.68	0.82	0.71	0.69
	7b – 0.01 WN	0.66	0.66	0.66	0.66
	7b – 0.05 WN	0.62	0.79	0.63	0.57
	7b – 0.1 WN	0.60	0.60	0.60	0.60
	7b – 0.2 WN	0.58	0.77	0.58	0.58
	7b – 0.3 WN	0.59	0.77	0.59	0.5
	7b – 0.4 WN	0.59	0.77	0.59	0.5
	7b – 0.5 WN	0.57	0.77	0.58	0.48

Table 6: CNN denoising defensive tests

Finally, the new SNR-based experiments were held. While these were not large-scale experiments, they nonetheless show that is possible to use the SNR-based function to successfully generate disturbed samples that adhere to the SNR standard, and subsequently to apply these samples into an ample variety of adversarial attacks.

Regarding stealthiness, even though one can physically perceive a given noise as being less loud (and hence stealthier) than the others, in practice it was not possible to pinpoint a given type of sound as being better suited for a practical adversarial attack. Although the experiments held were done in lab-level fashion, the digitally disturbed samples are enough to establish that even if these experiments were to be held on the field through loudspeakers, these different types of noise alone would not make the attacks stealthier in any significant way. This could be different if directional speakers were being used.

All the reported SNR experiments were based on the gunshot audio class.

Noise	**SNR threshold**	**Total # of samples**	**# Of successful detections**	**# Of failed detections**
Blue	10	50	7	43
Blue	20	50	29	21
Blue	30	50	46	4
Pink	10	50	9	41
Pink	20	50	33	17
Pink	30	50	47	3
Brown	10	50	9	41
Brown	20	50	34	16
Brow	30	50	46	4

Table 7: SNR-based Experiments

3.6 Related Work

3.6.1 Adversarial Attacks (on Speech Recognition Systems)

Personal assistants and speaker identification systems have become part of our daily lives. Recently, a large body of research has focused on studying the robustness of speech recognition systems against different types of adversarial attacks [Schonherr2018, Li2020]. For instance, the work by [Li2020] distinguishes itself on this front, as unlike the others, it does not require the attacker to know the original voice command in advance before attacking it and modifying it to make it malicious.

3.6.2 Adversarial Attacks on AED Systems

Much less work exists on this front. Subramanian et al. [Subramanian2020] studies attacks done against the audio portion of audio tagging systems, exploring its transferability properties across different deep learning models. [Subramanian2020] also shows that such transferability of adversarial examples can resist normalization techniques as well as knowledge distillation defense. Such attacks are not easy to reproduce in a real-world scenario, as they require some costly computations plus some technical savviness by the attacker. Besides, it is not clear how perceivable such disturbances are.

3.6.3 Countermeasures Against Evasion Attacks

Some work has studied countermeasure techniques for improving the resilience of these system against adversarial attacks [Roy2018, Carlini2018, Mao2020]. Most of these techniques are passive in nature, such as on the case of promoting the detection of an adversarial attack occurrence. Active techniques, such as adversarial training exist and can also be found in smaller numbers.

Active techniques, such as adversarial training exist and can also be found in smaller numbers. For example, Sallo et al. [Sallo2020], used six different attacks, all tampering the spectrograms (images) and not the audio files, employing them next against some publicly AED available models. The adversarial training on this case consists of using adversarial spectrograms.

3.6.4 Neural Network Approaches for AED

The works by [Zhou2017] and [Jaiswal2018] use a combination of Convolutional Neural Networks (CNN) with sequential layers and spectrograms for sound detection and classification, the first targeting urban sound (as air conditioners, jackhammers, etc.) and the second targeting deforestation sounds. Both authors' implementations are done through Keras python library and achieve accuracies between 40 and 85 percent for different datasets under use.

[Khamparia2019] uses two parallel rather than sequential hidden layers for sound classification. The spectrograms are generated through Matlab. 10 ambient sound classes (rain baby cries, sneezing, etc.) are used, and the experiments

show that the proposed model achieves nearly 78 and 50 percent classification accuracies for ESC-10 and ESC-50 respectively.

[Li2016] focuses on surveillance related sounds, and its main contribution is to use, rather than the spectrograms themselves, a combination of partitioned monochrome images derived from spectrograms. Such derivation is obtained through the application of Gabor filters to the original spectrograms, and these derivate images are the ones to be fed to a K-Nearest-Neighbor classifier. Such approach is claimed to achieve an average of 96 and 83 percent performance in terms of classification accuracy.

Both [Lim2017] and [Cakir2017] use Recurrent Neural Networks (RNN) and seek to classify suspicious events. The model by [Lim2017] uses a CNN first and its output is further fed to a RNN (two models in tandem), while [Cakir2017] uses recurrent layers and standard CNN layers in an interleaved fashion (one single model). Both authors address the vanishing Gradient problem differently, [Lim2017] using Long Short Term Memory Unit (LSTM) while [Cakir2017] using Gated Recurrent Unit (GRU). The two authors claim their approaches slightly outperform works based solely on CNNs.

An ensemble of CNNs is used by [Lee2017] to perform urban sound classification. Two independent models take spectrograms as inputs and compute individual predictions, while a final prediction is obtained by assembling both models' probabilities. The proposed model achieved 0.536 in the event-based F1-score and 0.66 in the segment-based error rate in evaluation set of DCASE2017. [Ghaffarzadegan2017] uses an ensemble of Deep CNN, Dilatated CNN (DCNN) and Deep RNN for rare events classification. LSTM is used and an average F-Score of 91.2 percent is achieved.

3.7 Chapter Conclusion

The main contributions of this doctoral research at phase 2 are two-fold: first, the results more definitively confirm that AED systems are vulnerable to evasion attacks by adversarial examples made of audio samples. AED-capable CNNs as well as third-party devices were tested, and while their initial baseline performance was good under ideal circumstances with regards to audio event detection, significant drops in classification performance were witnessed, when either background noise or white noise were injected into the audio samples.

Another important contribution was to shed clear light over the fact that not all types of noise are effective in decreasing the performance of classifiers. For example, while white noise infused to gunshot samples can significantly decrease the performance of gunshot detection classifier, adding white noise to glass break samples show much smaller decreases. While the attack approaches were shown to be effective, the defense ones used against the adversarial examples were also shown to be good.

Also important was to establish that different noises (white, pink, brown) are not stealthier among themselves and by themselves in any significant way. This is a limitation if these noises, in pure form, are used as part of a practical on-the-field attack. Under this limited scenario, white and background noise seem to be stealthiest noise possible due to their nature of being regularly present around people, All noises could though be used in stealthy fashion if one considers Signal-To-Noise-Ratio thresholds.

For instance, employing adversarial training leads to significant improvements. The potential of spectral gating denoising techniques was also verified, which when applied to test sets, led to better classification performance. As previously stated, this research is done under the motivation of being one step ahead of a future where Audio Event Detection Systems are going to become ubiquitous, hence being employed not only at homes, but also public spaces.

As such, it is important to motivate researchers from the academy and professionals from the industry to think of potential security shortfalls before executing the design and the implementation of AED solutions, thus paving the way for a safer and more effective future. Further tests are conducted during phase 3 to validate, once and for all, the conclusions reached as part of phase 2. A final focus on the stealthiness of the attacks will also be part of phase 3.

CHAPTER 4: ROBUST ATTACKS AGAINST AED SYSTEMS

In this chapter, the final phase of this research, namely phase 3, is presented. Chapter 4 structure generally follows the one introduced previously in Chapters 2, and 3, hence an introduction is provided in section 4.1; the specific research questions being addressed in this phase are brought in section 4.2; the studies proposed to answer these questions are described in section 4.3; the final research experiments are presented and discussed in section 4.4. Finally, chapter 5 brings up this doctoral dissertation final conclusions and contributions.

4.1 INTRODUCTION

The third and last phase of this research continues to work on the previously introduced audio-based disturbances, however, expands them further through the development and testing of new versions of these attacks, now with a renewed focus on stealthiness, thus, on making these attacks even more practical, feasible, malicious, and more disruption-capable in a real-world scenario.

Besides the intent to make the attacks stealthy, the renewed focus on novel field experiments is justified after there were several software updates released by Google for Nest devices over the course of the 12 months that passed since Phase 2 experiments were completed. These updates could have improved the AED capabilities of these Nest devices. Also, important to note is the availability of other AED-capable devices, such as Alexa ones, manufactured by Amazon. These are also popular equipment that are constantly being improved by their manufacturer.

As such, expanded on-the-field experiments, now employing both the baseline (noisy) as well as the new stealthy attack variants developed for phase 3, are conducted against an also expanded roster of AED devices, thus allowing for current capability evolution assessment, besides enabling comparisons of different detection performances provided by different devices of different manufacturers.

Given the focus on stealthiness, directional speakers are employed to fulfill the purpose of less perceptible attacks. These stealthy experiments were expected to increase the success rate of the adversarial attacks, while still maintaining simplicity through the adoption of only a few tweaks when compared to the conventional, loudspeaker-based attack variant. In other words, maximum discretion is to be prioritized as an important research design constraint.

Finally, once back to the lab environment, an end-to-end (E2E) AED system is proposed, where we couple our in-house built classifiers to an input capturing embedded microphone. The audio input provided was made of digital samples used on previous experiments as well as of audio captured directly by the black-box AED devices (hence that passed through their entire pipeline), thus representing, and evaluating an actual AED system to the largest extent possible.

4.2 Background and motivation

Since Audio Event Detection (AED) Systems have left the realm of theory and became a practical reality, the variety of AED designs is constantly increasing. From somewhat rudimentary, open designs based on low-cost platforms such as raspberry pi computers, to fully black-box, proprietary, and state-of-the-art

systems, these devices are generally expected by its users to provide reliable detection capabilities.

With several of these devices being marketed by its manufacturers as being intended for use as part of broader safety / security-driven frameworks, it becomes imperative to evaluate and to compare the reliability in terms of their detection performance, especially when mainstream brands are manufacturing these devices. During phase 2 of this research, attention was provided to Nest devices, manufactured by one of the top players in the industry, namely Google.

The results of the tests with the Nest devices were not particularly favorable, which inevitably led to questions about if the conducted tests were properly carried out and if its results could be trusted. This was a clear limitation from phase 2. Yet another limitation was the fact that only audible disturbances were employed as part of attacks against these Nest devices. While these noisy disturbances were somewhat common, made of day-to-day noises that could deceive nearby standers to some extent, they could not be considered sophisticated enough.

Phase 3 addresses all these shortfalls, first by adding to the field tests some Echo devices, manufactured by Amazon, and compatible with Alexa voice assistant services. The addition of another widely known mainstream brand to the tests allows for a chance to further validate the previously obtained results during Phase 2 tests. It also allows for clear and straightforward comparison of detection performance between these two major players, namely Google and Amazon.

Another contribution to be derived from Phase 3 resides on expanding the noisy attack envelope, now by adding to it some stealthy disturbances. While Phase 3 audio disturbances will still rely on the same backbones from Phases 1 and 2, namely white and background noises, and continue to be injected to the

environment through conventional loudspeakers (hence through audible, non-stealthy means), they will now also be introduced by directional speakers.

The use of directional speakers for the injection of audio adversarial examples is a major leap forward in terms of this research. This is because the proposed attacks remain practical to be reproduced on the field, especially by not so much savvy attackers (thus not breaking one of the research design constraints), but now also leads to greatly reduced chances of having the attack detected and perceived by nearby standers. In other words, the noisy attacks become stealthy, thus greatly increasing attack effectiveness and its chances of causing harm and of being disruptive to AED capable devices.

Given that both Nest and Alexa devices come from the factory being capable of detecting glass break sounds, said sound class will be phase 3 positive class (the one containing the sounds of interest). In the revised threat model for phase 2 and evolved for phase 3, the AED system will remain as a simulation of a deployed home security system, and the adversary will keep aiming at preventing the AED system from correctly detecting and classifying the sound events.

For said AED defeating purpose, the adversary will still generate noise (background or white) through audible means, but now will also add to his arsenal stealthy noise. thanks to the adoption of directional speakers.

4.3 Threat Model

Phase 3 threat model is a direct evolution from the one proposed for phase 2. As such, in it loudspeakers and directional speakers alike will be employed, having both to reproduce disturbances comprised of white and background noise.

These two types of noise were carefully selected to maximize imperceptibility in all scenarios, as even when audible and if heard, these two sounds can easily be mistaken for ordinary noise, possibly found within almost any environment.

In other words, for the case of loudspeakers, while even though both noises would be able to be perceived by local bystanders due to how loudspeakers work, it is unlikely these sounds would draw too much attention or concerns, as they could be mistaken by mere environmental noise. Even pure white noise would be much less conspicuous than, for instance, audible and clearly stated voice commands as it would be the case of AEs employed against SR systems.

On the other hand, when employing directional speakers, the perceptibility of the disturbances is strongly reduced, once again, due to the very way directional speakers work. Unlike conventional speakers, which spread their sound waves through a wide area, directional speakers work as a flashlight instead [ExplainThatStuff2020], except that they use audio rather than light.

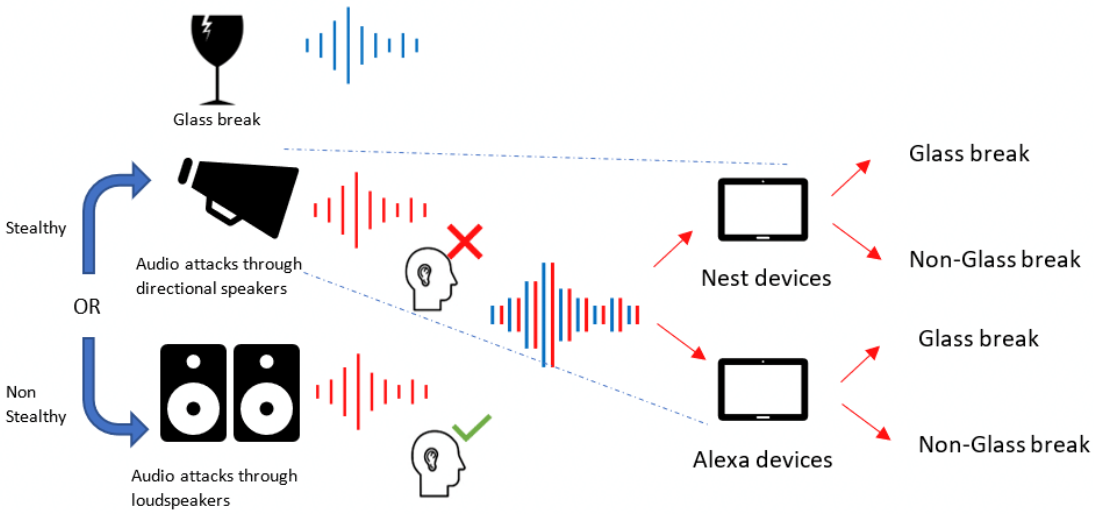

Figure 9: Updated threat model with stealthy disturbances

This peculiar way to work allows for an attacker to use directional speakers to single out a particular targeted AED device, directly focusing against it an audio attacking beam, thus not affecting (or at least affecting very little) whoever may be standing nearby the spot of attack. As part of phase 3 tests, both directional speakers (used for novel, and stealthy attacks), and loudspeakers (for re-executing and validating results from conventional attacks) are employed.

Figure 9 brings a representation of the previously described threat model, where it is assumed that the adversary actively attempts to evade an AED system that aims on detecting suspicious sound events. A black-box scenario also remains as an assumption, so once again the adversary will not have any knowledge about the datasets, algorithms, and their parameters.

4.4 Materials and Methods

4.4.1 Third-party AED Devices

The same as it happened during phase 2, the focus of testing AED capability effectiveness (or lack thereof) remains directed against mainstream, widely known AED devices that are also available for purchase by the public. Since there was the need to re-evaluate and to confirm the results from phase 2 field experiments, it is just natural that Nest devices would be included for testing under phase 3. Both Nest display and minis are reincluded, as such.

As previously explained, there was the need to include into testing new devices belonging to different brands. Echo devices, manufactured by Amazon, powered by Alexa assistant services, were chosen for this purpose. Being equivalent in capability to the selected Nest devices, the Echo display and the

Echo dots offer a good alternative for Nest devices and having both brands included into Phase 3 field tests is a good way to employ state-of-the-art devices.

4.4.2 End-to-End AED System

Considering all the components that are needed to put together a modern AED system / device, similarly to what is found on Nest and Alexa devices, it made sense to reproduce, at least partially, a full E2E AED system for testing purposes. For that, the glass break classifier, trained on unnoisy samples and that has been extensively experimented during Chapter 3, was coupled with a microphone for audio capture / feeding / classification purposes.

To simulate the embedded microphones found on actual AED devices, the embedded microphones from a MacBook Pro 2021 computer have been used, as they are similar in size and capabilities. Despite the similarities, this simulated AED system cannot possibly fully reproduce the entire data pipeline used in actual AED black-box devices, as it misses important components (such as embedded machine learning chips as well as cloud-based services), many of these components being unknown (hence black-box).

To partially address this concern, in a second moment, audio captured by Alexa devices over the course of the field experiments (and that as such, are guaranteed to have passed through the entire pipeline before being stored) has been retrieved and will be used as part of the experiments. This is because Alexa devices only store and make available for download audio that has been successfully detected as being an occurrence of the audio event of interest, on this case, glass break. If undetected, the audio is not stored, hence is not available.

4.4.3 EXPERIMENTS

As previously stated, the main objective by Phase 3 was to confirm the results reached as part of phase 2 while also giving focus on the stealthiness aspect of the noisy attacks being employed. For that purpose, a large field experiment consisting of several smaller experiments was devised. These were experiments 8a to 8y, where different combinations of Nest and Echo devices were tested.

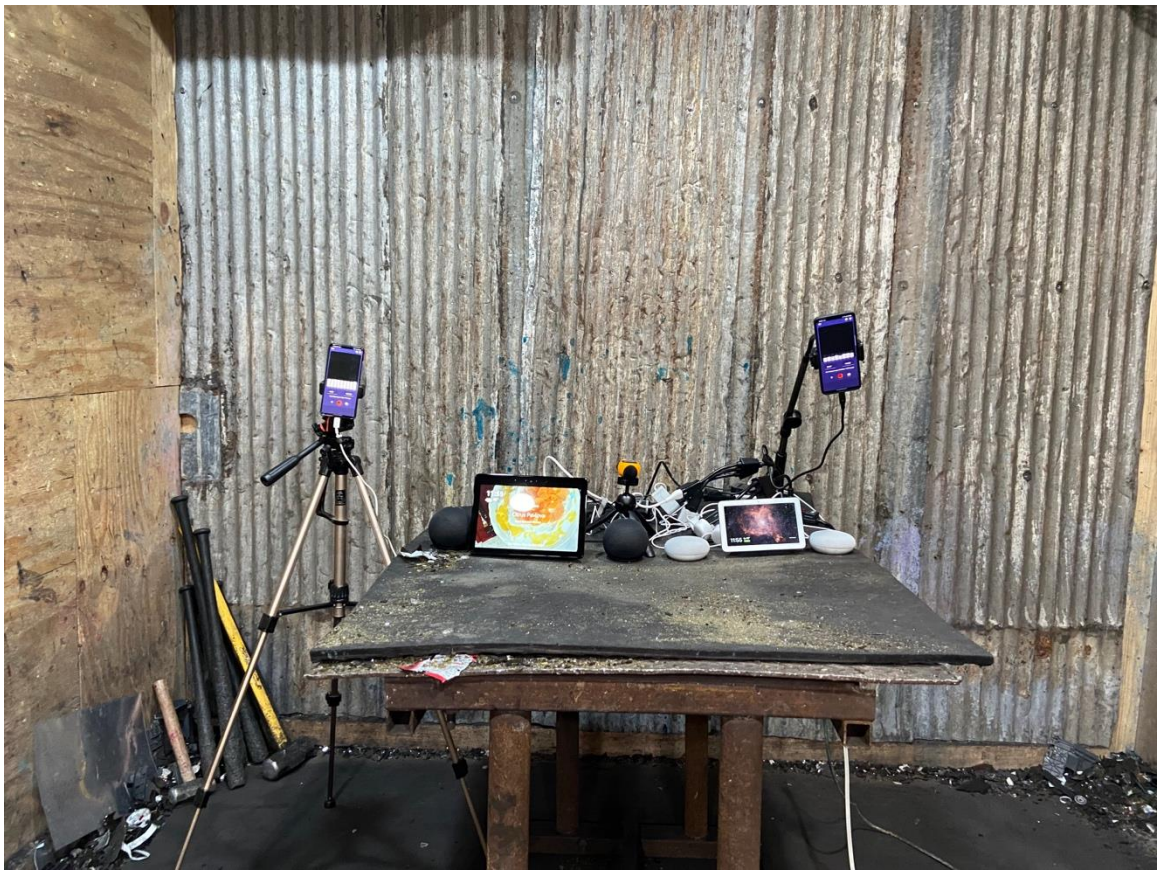

Figure 10: Late field experiments

By employing each device in isolation from each other or working in some sort of combination (for instance, two nest minis working together) the idea was to

assess if more devices made any difference in how good the actual AED capabilities by each brand was. It was also sought to identify how distance from the glass break audio source to the AED devices would affect the results. Impacts from (bad) internet connection on the experiments were ruled out by having all devices connected to a high-speed 5G network.

For real glass break sounds, previously purchased beer bottles were broken. To record the whole procedure, but also to generate some glass break sounds for possible later reuse, two Android devices were employed, namely S20 Ultra and S21 Ultra, both working as audio recorders, positioned at negligible distance from the AED devices.

Figure 11: Loudspeakers (table) and directional speaker (tripod)

To establish signal-to-noise ratio readings, the room where the experiments were conducted had its environmental noise measured when being free of any experiment-related sound, baselining at 40 decibels by then. Two loudspeakers were deployed, namely Charge 4 and Flip 4, both manufactured by JBL, the former as a non-stealthy attacking device and the latter to reproduce digital glass break sounds. An FS-mini-B directional speaker manufactured by VidBeam, playing the role of stealthy attacking device was also employed.

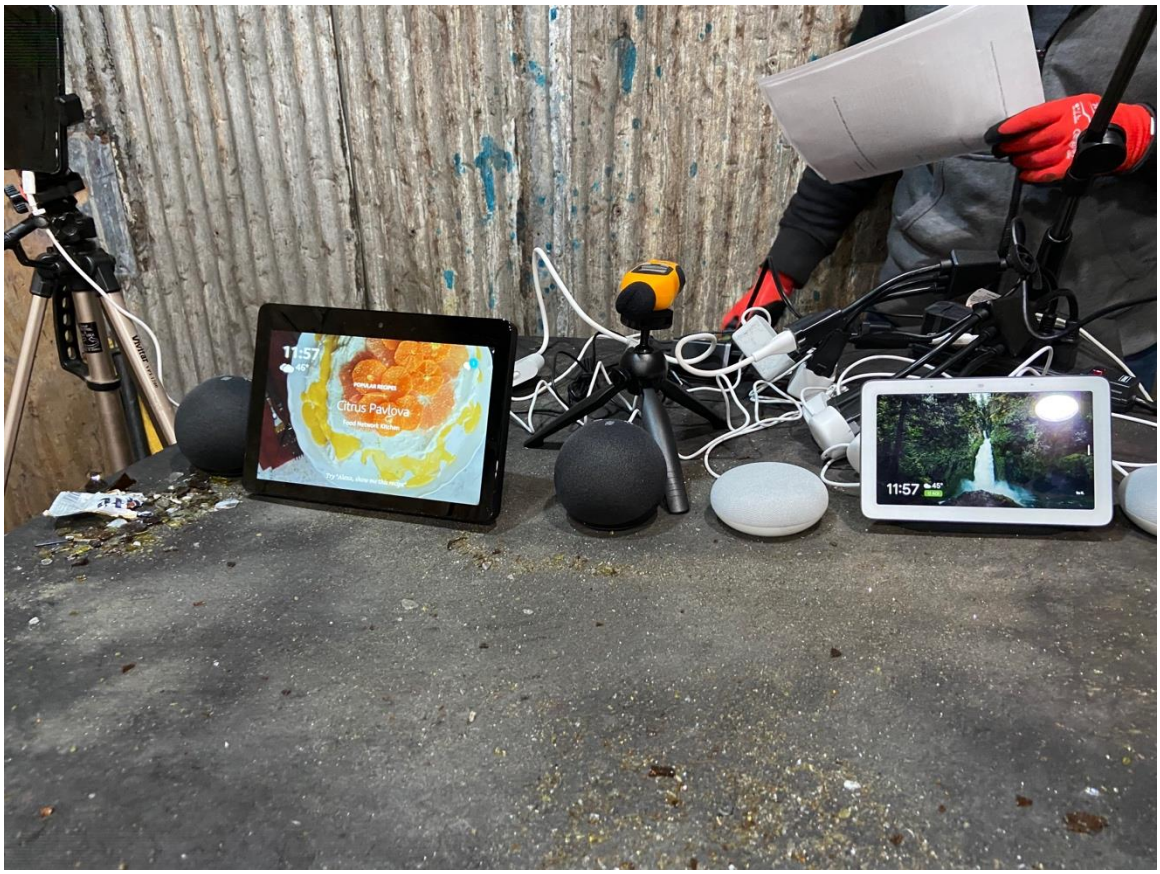

Figure 12: AED devices and decibel reader

All speakers were wireless connected, via Bluetooth, to one commodity laptop computer, where the audio disturbances as well as the digital glass break sounds were stored. Both loudspeakers had native Bluetooth connection, however

the directional speaker did not offer that capability. This roadblock was bypassed by coupling it to an A3352 Bluetooth Receiver by Anker, which effectively made the directional speaker wireless. The computer sound volume was set to 100 percent, while the loudspeakers had their volume set at 90 percent to ensure a Signal-to-Noise (SNR) ratio of 30 decibels. The directional speaker had its volume set to 100, regardless of SNR achieved by doing so.

4.5 RESULTS

In total 254 real glass breakages were performed to test the standard detection capabilities from the AED devices (when detecting sounds under normal conditions of operation) as well as when these capabilities were supposed to be impaired by adversarial noisy attacks. As such, out of these breakages, 54 happened when all devices were detecting sounds under no attack.

Another 120 happened under non-stealthy attacks (white and background noises injected into the environment by a loudspeaker), and finally the remaining 40 were done under stealthy attacks (noise beam from the directional speaker targeting a single active detecting device at a time). Another 150 digital glass breaks (playing glass break sounds through a loudspeaker) were performed, out of which 70 happened under no attack, 40 happened under noisy attacks, and finally the remaining 40 happened under stealthy attacks.

As it can be seen from Figure 14, only 23 of the real glass break occurrences (5 by both Nest and Alexa devices at the same time, 8 by Nest devices alone and 10 by Alexa devices alone) and 38 of the digital ones (18 by both Nest

and Alexa devices at the same time, 11 by Nest devices alone and 9 by Alexa devices alone) have been detected.

From these extremely low number of successful glass break detections, one can deduct that the detection capabilities of both Nest and Alexa devices are not good to start with, even when they are operating under normal condition of use, meaning facing no attacks and when signal to noise ratio favors detection by being above 30 decibels. Keeping SNR above 30 is, by the way, the reason why some of the real glass experiments were done by breaking two pieces of glass at the same time.

The noisy attacks, audible and stealthy alike, only make the poor performance presented by these AED devices even worse, reason why an attacker could make use of them to almost guarantee one hundred percent success rate while attacking AED devices. The tests prior to the disturbances already proved that an adversary is very likely to be successful in breaking glass (like a glass window) and still avoid detection while doing so.

In fact, the performance by the AED devices, Nest and Alexa alike, is so poor that reporting the distance factor (between detecting devices and breaking spot) becomes irrelevant: while the intention was to break glass at 2 and 4 meters away from the detecting devices, it was quickly realized that all devices performed very poorly, even when only 2 meters away from the breakage spot.

Test	Experiment Id.	# Of samples	# Of detections	Detecting devices	Observations
Initial Setup	8a – Real Glass – Displays	3	0	N/A	
	8b – Real Glass – Minis + Dots	3	0	N/A	
	8c – Real Glass – All	2	0	N/A	
	8d – Digital – Displays	10	0	2 by both + 2 by Nest	
	8e – Digital – Minis + Dots	10	0	1 by both + 1 by Alexa	
Baseline	8f – Real Glass – All	10	0	N/A	
	8g – Real Glass – Displays	20	7	3 by Nest, 3 by Alexa, 1 by both	Two glasses for each breakage
	8h – Real Glass – Minis + Dots	20	5	1 by Nest, 3 by Alexa, 1 by both	Two glasses for each breakage

	8i – Real – All	20	4	1 by Nest, 3 by both	Two glasses for each breakage
	8j – Digital – Displays	20	8	3 by Nest, 5 by both	
Noisy	8k – Digital 0 Minis + Dots	20	11	2 by Nest, 3 by Alexa, 6 by both	
	8l – Digital – All	20	9	1 by Nest, 4 by Alexa, 4 by both	
	8m – Real Glass – N-WN- All	10	1	1 by Nest	Two glasses for each breakage
	8n – Real Glass – N-BN- All	10	1	1 by Nest	Two glasses for each breakage
	8o – Real Glass – N-WN - All	10	0	N/A	Two glasses for each breakage
	8p - Real Glass – N-BN - All	10	2	2 by Alexa	Two glasses for each breakage

	8q - Digital – N-WN - All	10	1	1 by Nest	
	8r - Digital – N-WN - All	10	2	1 by Nest, 1 by Alexa	
	8s - Digital – N-WN - All	10	1	1 by Nest	
	8t - Digital – N-BN - All	10	0	N/A	
Stealthy	8u - Real Glass – S-WN – Nest Display	20	1	1 by Nest	Two glasses for each breakage
	8v - Real Glass - S-WN – Echo Display	20	3	3 by Alexa	Two glasses for each breakage
	8w - Digital – S-WN – Nest Display	20	0	0	
	8y - Digital – S-WN – Echo Display	20	0	0	

Table 8: Compilation of late field tests (Nest and Echo)

		Non-Stealthy	Stealthy*
Total Digital Samples Available		140	10**
Total Real Samples Available		174	80**
# of Digital Detections	E	9	0
	N	11	0
# of Digital Misdetections	E	113	5
	N	111	5
# of Real Detections	E	10	3
	N	8	1
# of Real Misdetections	E	159	37
	N	161	39
# of Digital Detections	B	18	NA
# of Real Detections	B	5	NA
Total Real Digital Detected out of 254 (total when under attack, WN Attack, BN Attack)		27 (8, 5 WN, 3 BN) Echo 15, Nest 13	
Total Digital Real Detected out of 150 (total when under attack, WN Attack, BN Attack)		38 (4, 2 WN, 2 BN) Echo 27, Nest 29	
*: Only displays participated on stealthy experiments			
**: reported samples were split among displays			

Table 9: Summary of late field tests. Results reported by Echo(E), Nest(N) and both Echo and Nest (B) devices together.

As such, since no clearly distinguishable practical performance difference in terms of AED detection done between 2 and 4 meters could be found, all tests were conducted and reported with 2 meters between glass break spot and detecting devices. This is restriction of use that also makes the AED devices under testing not to be very practical for day-to-day use, as in a real situation they are supposed to be deployed into different environments under all sorts of different conditions, including distance and varying levels of environmental noise.

About the E2E AED system tests, the baseline 1a glass break classifier from chapter 3 (trained on 2000 samples, half being glass breaks) has been reused. The corresponding 1a test set, made of 300 test samples (150 of which were actual

glass break sounds) has also been reused, except that now, all these samples are played over the air by a Samsung S21 Ultra device, and captured by the MacBook Pro embedded microphones before being fed directly to the glass break classifier.

Test	Total Positive Samples	Total Negative Samples	Positive Classifif. (Correct)	Negative Classific. (Correct)	Positive Misclassif. (Incorrect)	Negative Misclassific. (Incorrect)	A C C
Digital Glass Break	150	150	74	144	76	6	0.726
Alexa Glass Break	35	NA	26	NA	9	NA	0.742

Table 10: E2E AED System Tests

The results of these tests can be seen on table 9. As it is clear, the addition of a microphone does adversely affect the classifier performance, as it drops from 99% accuracy (as reported during chapter 3) to a little less than 50%, as it misses 76 samples out of the 150 total. Still, this reduced performance is nowhere near the subpar results obtained from the black-box AED devices, being, as such, much better. The echo captured glass breaks samples were tested next.

From table 9 it is possible to see that a test set made of 35 samples were available, and out of these, only 3 misclassifications happened. From previous observation along the research, it was already clear that signal-to-noise ratio was a very important factor for any glass break to be detected by black-box AED

devices. The good performance of the in-house built classifiers on top of these Alexa captured samples is further evidence of SNR importance across the board.

No negative samples are available in the test set, as it is only possible to access and download sounds that were successfully detected, processed, and stored by the Alexa devices, which obviously does not apply to random sounds that do not contain glass break. This also late evidence that the overwhelming majority of the glass breaks from chapter 3 field experiments were really never neither detected, nor stored, at least by the Alexa devices.

4.6 Related Work

This section covers other authors' work related exclusively to phase 3. Its focus is on the several components making up physical speech recognition and / or audio event detection systems.

4.6.1 Components Making Up Physical Systems

For practical SR and AED systems to work properly (for either voice to be recognized or for acoustic events to be detected), it is known that several components working together are needed. For instance, while working on voice recognition, [Oh2019] implements a system made of a microphone for signal capture, followed by custom processing chips, one for local feature extraction, followed by another chip responsible for applying a locally deployed classifier to receive such features as input.

Several other components exist, such as post-processing chips, mixers, amplifiers, and others. Author [Li2021], while also working on voice/speech related-applications, proposes a real-time on-chip speech audio super resolution system, made of dual microphones (bone conduction as well as air conduction) at the edge of the system, followed by system on chip for input intake, GPUs for training the ATS-UNet deep learning model, which is to be deployed to off-the-shelf ARM-Cortex micro-controllers for on-device local processing.

Author [Alsina2017] implements a system for real-time AED for the support of Ambient Assisted Living. For that purpose, a wireless acoustic sensor network with several low-cost microphone nodes captures environmental sound and sends it to a GPU-based concentrator, that acts as an MFCC feature extractor that is passed along to a NVIDIA GPU for locally detecting the acoustic events of interest.

Recently, AI chip startup Aspinity [EETimes2021] made news when it released a specialized acoustic event detection chip that will most likely become an off-the-shelf component, able to be integrated into diverse battery-powered devices deployed as part of AED capable solutions.

4.7 Chapter Conclusion

During phase 4, an expanded roster of AED capable black-box devices was evaluated regarding both their AED performance as well as their vulnerabilities when facing evasion attacks based on noisy disturbances. The results obtained from these tests were clear to show that even modern devices, considered to be state-of-the-art, present a far than ideal performance when detecting sounds of

interest. This is especially troublesome given the safety / security scenarios where these devices tend to be employed as part of.

Several reasons may be behind this less than optimum performance, and it is likely that each component within the required data pipelines may is to blame, at least partially. Experiments held to simulate an end-to-end AED system and that coupled an audio input device to the in-house built classifiers, despite being simple, is already enough to bring a dramatic reduction in detection performance, generating a drop in accuracy from 99% to less than 50%. Despite such reduction, the proposed AED system significantly outperforms the state-of-the-art.

These results seem to demonstrate that the current limitations of these AED devices are most likely not tied to the deep learning models / data in use, but to too many components being part of the pipelines, which may be there to support other capabilities / objectives rather than sound event detection in itself. These other objectives may be tied to things as security, personalization, localization services, besides others, and may be a good example of secondary capabilities adversely impacting the primary ones.

CHAPTER 5: CONCLUSIONS

It is undeniable that Audio Event Detection is a capability that has gained tremendous improvements in performance over the last decade. Such improvements are one of the reasons that industry juggernauts, such as Google and Amazon (just to cite a few) to quickly embrace such capabilities and to adopt them as part of their industry-leading devices. Unfortunately, the adoption of AED capabilities by industry heavy weights may be deceiving.

This is because consumers may be led to wrongly believe that these are well-established devices with well-established capabilities, both of which are failure-proof, an especially dangerous assumption given their use for safety purposes. Also, the wide AED capability adoption by manufacturers, tied to the today's ubiquitousness of AED-capable devices may attract attackers who seek to find vulnerabilities within these systems and to exploit them for malicious purposes.

The main contributions of this dissertation are four-fold: first, this research tested and unequivocally proved that black-box devices that are AED capable, even when manufactured by major brands such as Google and Amazon, are not to be trusted in terms of their detection capabilities. These devices fail to detect most audio events of interest, even when the conditions favor the precise detection (no attack being carried out, signal-to-noise ratio above 30 decibels).

Second, it was shown that if the AED performance by these devices was neither good nor reliable, it only gets worse when these AED physical devices are under attack. In the process of attacking the AED devices, this research employed both non-stealthy as well as stealthy disturbances made of noise, both which

further degraded the tested AED capabilities, to the point of rendering any device useless.

Third, special focus was given on tests that targeted solely CNN AED classifiers, as they are representative of neural networks commonly embedded into these AED black-box devices, hence are considered to be the main component in the AED pipeline. It was confirmed that it is possible to build AED classifiers that are very good to detect a given sound of interest (e.g.: glass break).

While these were good news, the same tests also have shown that these same classifiers are clearly vulnerable to evasion attacks by adversarial examples made of noise, vulnerability that is similar to the one found for the black-box AED devices previously tested. Significant drops in classification performance were witnessed, when either background noise or white noise were injected into the audio samples.

However, an important observation stemming from the experiments was that not all types of noise were effective in decreasing the performance of classifiers in the exact same way or proportion. For example, while white noise infused to gunshot samples can significantly decrease the performance of gunshot detection classifier, background noise is not that effective against this class.

It was also shown that defenses against these noisy adversarial examples perform well when shoring up the affected classifiers. For instance, employing adversarial training leads to significant improvements. A denoising technique based on spectral gating has also shown to be effective, as it led to better classification performance.

Finally, an end-to-end AED system was proposed to showcase, at least partially, how these AED systems are impacted when new components are added

to their data pipelines. The addition of a single component was already able to cause a reduction in performance of nearly 50%.

5.1 Summary of Contributions

- Evaluation of the baseline performance of actual, state-of-the-art, black-box, physical AED-capable devices. Such baseline performance has been decisively shown to be poor to say the least, being unacceptable when considering deploying such devices as part of any serious safety-driven framework.
- Evaluation of these same physical AED-capable devices against both audible and stealthy real disturbances, crafted and employed on-the-field. These disturbances have been shown to be able to render the state-of-the-art AED devices effectively useless.
- Focus on the implementation and evaluation of the single and most critical component making up modern AED capable devices, namely AED specialized neural networks. A reasonable baseline performance for these neural networks, much superior to the one found when testing the whole black-box AED devices, was established.
- Evaluation of AED neural networks against attacks made of adversarial example inputs made of noise. These adversarial examples were crafted to reproduce the physical attacks conducted against the AED

devices. It was shown that these neural networks are fundamentally vulnerable to these adversarial examples.

- Adversarial training and denoising were shown to be able to be used as countermeasures to the audio adversarial examples employed against the neural network AED component. While it is not a full solution regarding the performance of the entire AED system, this strategy can be used to improve the performance and robustness of at least the neural networks that make up the core of modern AED devices.

- Proposed an end-to-end AED system, made of in-house built classifiers, coupled with embedded microphone component. The proposed AED system performs better then mainstream AED-capable devices, such as Nest and Echo ones. The same system is an important contribution to demonstrate how the addition of components into the data pipeline currently tends to generate worst classification results, as a single component led to over 50% drops in accuracy. This may provide a useful insight for AED system designers.

5.2 Future Work

Future research spawning from this work may include:

- New stealthy noisy attacks derived from algorithmic means, more specifically on a modified, gradient-based version of the attack successfully used so far. While several approaches for gradient-based attacks exist, the Projected Gradient Descent (PGD) method seems to

be ideal, as it was already conceived to be somewhat robust and hard to be detected, as it uses small step sizes and generates small but progressive perturbations that are effective and fast to generate. Gradient based attacks denote white box access to the model under attack so that the gradients are known, something never considered within this doctoral research threat models. As such, it is important to clarify that the approach to be followed will be to use a surrogate model, simple in nature, from which the perturbations will be derived, and then, by exploring the well-known adversarial example transferability property, will be applied to the final AED model.

- Stealthy noise attacks carried by a mobile vector, such as drones, could also be researched. Considering that attacks based on directional speakers were tried successfully during the last phase of this research, it may be a good direction to turn these attacks into an even stealthier through drones, as they not would not only be hard to hear, but also hard to see.
- To research novel and effective defenses against the noisy disturbances. These defenses could maintain their reliance on the combination of oversampling, adversarial training and denoising techniques. For adversarial training, a logical evolution would be to include PGD on it. For denoising, further improvements to the spectral gating denoising technique might be pursued together with other forms of denoising. Another possibility under consideration is the employment of compressed deep learning models, an approach shown to be more robust to attacks on the speech recognition domain.
- Another advance that may be pursued is the research of defenses that could immediately and straightforwardly be employed on the field. Given the black-box nature of the AED capable devices that are part of this study, hence, the lack of access to their internal structure, such defenses may not necessarily be algorithmic, but may rely on hardware approaches instead. One possibility is the addition of proportional additive gaussian noise to the environment, hence attenuating maliciously injected noise.

REFERENCES

[Abdullah2019] Abdullah, H. and Garcia, W. and Peters, C. and Traynor, P. and Butler, K. and Wilson, J. Practical hidden voice attacks against speech and speaker recognition systems. IEEE Symposium on Security and Privacy, 2019.

[Ahsan2016] Ahsan, U. and Bais, A. “A Review on Big Data Analysis and Internet of Things”. 13th International Conference on Mobile Ad Hoc and Sensor Systems, 2016.

[AirborneSound2017] AirborneSound. The Free Firearm Library - Expanded Edition. Available at https://www.airbornesound.com. Last accessed in May 2020.

[Akhtar2018] Akhtar, N. and Mian, A. Threat of adversarial attacks on deep learning in computer vision: A survey. IEEE Access, pp. 14410-14430, 2018.

[Alexa2019] CNBC. How to Set up Alexa Guard on an Amazon Echo. Available at www.cnbc.com/2019/05/14/how-to-set-up-alexa-guard-on-an-amazon-echo.html. Last accessed in January 2021.

[Alraddadi2019] Alraddadi, S. and Alqurashi, F. and Tsaramirsis, G. and Luhaybi, A. and Buhari, S. Aroma Release of Olfactory Displays Based on Audio-Visual Content. Appl. Sci., pp. 1919-1958, 2019.

[Alsina2017] Alsina-Pagès, R.M. and Navarro, J. and Alías, F. and Hervás, M. homeSound: Real-Time Audio Event Detection Based on High Performance Computing for Behaviour and Surveillance Remote Monitoring. Sensors 2017, 17, 854.

[Al-Fuqaha15] Al-Fuqaha, A. and Guizani, M. and Mohammadi, M. and Aledhari, M. “Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications”. IEEE Communication Surveys & Tutorials, pp. 2347-2376, 2015.

[Athalyde2017] Athalye, A. and Engstrom, L. and Ilyas, A. and Kwok, K. Synthesizing robust adversarial examples. arXiv preprint arXiv:1707.07397, 2017.

[Analyticsindiamag2019] Analystics India Magazine. How Machine Learning Rocked Google's Hardware Event This Year. Available at https://policyadvice.net/insurance/insights/google-home-statistics. Last accessed on December 2020.

[Audacity2021] Audacity. How Audacity Noise Reduction Works. Available at www.wiki.audacityteam.org/wiki/How_Audacity_Noise_Reduction_Works. Last accessed on May 2021.

[Austin2020] ServiceASAP. Comprehensive Gunshot Detection Systems. Available at: https://www.serviceasap.com/solutions/commercial-solutions/gun-shot-detection. Last accessed in May 2020.

[Bhattacharya2020] Battacharya, S. and Manousakas, D. and Ramos, A. and Venieis, S. “Countering Acoustic Adversarial Attacks in Microphone-equipped Smart Home Devices”, Interactive, Mobile, Wearable and Ubiquitous Technologies, Volume 4, Issue 2, 2020.

[Bilen2020] Bilen, Ç.; Ferroni, G.; Tuveri, F.; Azcarreta, J. A Framework for the Robust Evaluation of Sound Event Detection. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4-8, 2020.

[Boyat2015] Boyat, A. and Kumar and Joshi, B. A review paper: noise models in digital image processing. ArXiv preprint 1505.03489, 2015.

[Buda2018] Buda, M. and Maki, A. and Mazurowski, M. A systematic study of the class imbalance problem in convolutional neural networks. Neural Networksh, pp. 249-259, 2018.

[Busse2019] Busse, C.and Krause, T. and Ostermann, J. and Bitzer, J. “Improved Gunshot Classification by Using Artificial Data”, AES International Conference on Audio Forensics, 2019.

[Cakir2017] Çakır, E. and Parascandolo, G. and Heittola, T. and Huttunen, H. and Virtanen, T. “Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection”, IEEE/ACM Transactions on Audio, Speech, and Language Processing, pp. 1291-1303, 2017.

[Carlini2017] Carlini, N., Wagner, D. Towards evaluating the robustness of neural networks. IEEE symposium on security and privacy, pp. 39-57, 2017.

[Carlini2018] Carlini, N. and Wagner, D. Audio adversarial examples: Targeted attacks on speech-to-text. IEEE Security and Privacy Workshops, pp. 1-7, 2018.

[Chavez2017] CNN. Using Sound to Attack: The Diverse World of Acoustic Devices. Available at www.cnn.com/2017/08/10/health/acoustic-weapons-explainer/index.html. Last accessed on January 1st 2021.

[Clavel2005] Clavel, C. and Ehrette, T. and Richard, G. “Events Detection for an Audio-Based Surveillance System”, IEEE International Conference on Multimedia and Expo, pp. 1306-1309, 2005.

[Chiang2020] Chiang, P. and Geiping, J. and Goldblum, M. and Goldstein, T. and Nj, R. and Reich, S. “Witchcraft: Efficient PGD Attacks with Random Step Size”, IEEE International Conference on Acoustics, Speech and Signal Processing, 2020.

[Choi2005] Choi, Y. and Kim, K. and Jung, J. and Chun, S. and Park, K. Acoustic intruder detection system for home security. IEEE Transactions on Consumer Electronics, pp. 130-138, 2005.

[Chu2004] Chu, W. and Cheng, W. and Wu, J. and Hsu, J. “A study of semantic context detection by using SVM and GMM approaches”, IEEE International Conference on Multimedia and Expo., pp.1591-1594, 2004.

[Cowling2003] Cowling, M. Comparison of techniques for environmental sound recognition. Pattern Recognition Letters, 2895-2907, 2003.

[Crocco2016] Crocco, M. and Cristani, M. and Trucco, A. and Murino, V. Audio Surveillance: A Systematic Review. ACM Computing Surveys, 2016.

[Dahlan2018] Dahlan, R. AdaBoost Noise Estimator for Subspace based Speech Enhancement. International Conference on Computer, Control, Informatics and its Applications (IC3INA), pp. 1-2, 2018.

[DCASE2017] DCASE. Detection of rare sound events. Available at www.cs.tut.fi/sgn/arg/dcase2017. Last Accessed in May 2020.

[Dennis2011] Dennis, J. And Tran, H. and Li, H. Spectrogram Image Feature for Sound Event Classification in Mismatched Conditions. IEEE Signal Processing Letters, pp. 130-133, 2011.

[Donmoon2017] Donmoon, L., and Lee, S. and Han, Y. and Lee, K. “Ensemble of Convolutional Neural Networks for Weakly-Supervised Sound Event Detection using Multiple Scale Input”, 2017.

[Dormehi2018] Digital Trends. U.S. Military Is Developing a Sound Weapon that Sounds Like a Retro Modem. Available at digitaltrends.com/cool-tech/military-sound-weapon-old-modem/. Last accessed on January 1st 2021.

[Dufaux2000] Dufaux, A. and Besacier, L. and Ansorge, M. and Pellandini, F. “Automatic sound detection and recognition for noisy environment”, 10th European Signal Processing Conference, pp. 1-4, 2000.

[Eagle2020] EagleTechnology. Comprehensive Gunshot Detection Systems. Available at https://www.serviceasap.com/solutions/commercial-solutions/gun-shot-detection. Last accessed in May 2020.

[Edmonds2006] Edmonds, E. Abstraction and interaction An art system for white noise. International Conference on Computer Graphics, Imaging and Visualisation (CGIV), pp. 26-28, 2006.

[ESC502021] Piczak, Carol. ESC-50: Dataset for Environmental Sound Classification. Available at https://github.com/karolpiczak/ESC-50. Last accessed on May 2021.

[EETIMES2021] EETimes. Aspinity Expands into Audio Event Detection. Available at https://www.eetimes.com/aspinity-expands-into-audio-event-detection/#. Last Accessed in March 2022.

[Evangelopoulos2009] Evangelopoulos, G. and Zlatintsi, A. and Skoumas, G. and Rapantzikos, K. and Potamianos, A. and Maragos, P. and Avrithis, Y. “Video event detection and summarization using audio, visual and text saliency”. IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3553-3556, 2009.

[ExplainThatStuff2020] Explain that Stuff!. Directional loudspeakers. Available at https://www.explainthatstuff.com/directional-loudspeakers.html. Last Accessed in March 2022.

[Fesliyan2021] Fesliyan Studios. Fesliyan Studios Royalty Free Music. Available at https://www.fesliyanstudios.com/contact. Last accessed on May 2021.

[Freesound2021] Freesound. Freesound. Available at https://freesound.org/help/faq/. Last accessed on May 2021.

[Fu2017] Fu, X. and Ch'ng, E. and Aickelin, U. CRNN a joint neural network for redundancy detection. Proceedings of the IEEE International Conference on Smart Computing (SMARTCOMP), pp. 29-31, 2017.

[Galloway2018] Galloway, A. and Taylor, G. and Moussa, M. Predicting adversarial examples with high confidence. arXiv preprint arXiv:1802.04457, 2018.

[Gao2018] Gao, S. and Lin, B. and Wang, C. Share price trend prediction using CRNN with LSTM structure. Proceedings of the International Symposium on Computer, Consumer and Control (IS3C), pp. 6-8, 2018.

[Ghaffarzadegan2017] Ghaffarzadegan, S. and Salekin, A. and Ravichandran, A. and Das, S. and Feng, Z. Bosch Rare Sound Events Detection Systems for DCASE 2017, 2017.

[Goodfellow2014] Goodfellow, I. and Shlens, J. and Szegedy, C. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014.

[Hansheng2013] Hansheng, L. and Valdez, O. “Special Sound Detection for emergency phones”, 10th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), pp. 816-820, 2013.

[Hendrik2017] Hendrik, J. and Chaithanya, M. and Brox, T. and Fischer, V. Universal adversarial perturbations against semantic image segmentation. Proceedings of the IEEE International Conference on Computer Vision, pp. 2755-2764, 2017.

[HHS2014] U.S. Department of Health and Human Services. Incorporating Active Shooter Incident Planning into Health Care Facility Emergency Operations Plans. Available at https://www.phe.gov/Preparedness/planning/Documents/active-shooter-planning-eop2014.pdf. Last accessed in April 30, 2019.

[Hodgson2010] Hodgson. Understanding Records: A Field Guide to Recording Practice. Continuum International Publishing Group, 2010.

[Hognelig2015] Hognelid, P. and Kalling, T. “Internet of Things and Business Models”. IEEE 9th International Conference on Standardization and Innovation in Information Technology, 2015.

[Huq2020] Huq, A. and Pervin, T. “Analysis of Adversarial Attacks on Skin Cancer Recognition”. International Conference on Data Science and its Applications, 2020.

[Husamuddin2017] Husamuddin, M. and Qayyum, M. “Internet of Things: A study on security and privacy threats”. 2nd International Conference on Anti-Cyber Crimes, 2017.

[Isar2016] Isar, D. and Gajitzki, P. “Pink noise generation using wavelets”, 12th IEEE International Symposium on Electronics and Telecommunications (ISETC), pp. 261-264, 2016.

[Ito2016] Ito, A. “Recognition of sounds using square cauchy mixture distribution”. IEEE International Conference on Signal and Image Processing, pp. 726-730, 2016.

[Jaiswal2018] Jaiswal, K. and Patel, D. “Sound Classification Using Convolutional Neural Networks”, International Conference on Cloud Computing in Emerging Markets, 2018.

[Jose2020] Jose, C. and Mishchenko, Y. and Senechal, C. and Shah, A. and Escott, A. and Vitaladevuni, S. Accurate Detection of Wake Word Start and End Using a CNN. ArXiv abs/2008.03790, 2020.

[Kesslen2019] NBC News. Plug your Ears and Run': NYPD's Use of Sound Cannons Is Challenged in Federal Court. Available at www.nbcnews.com/news/us-news/plug-your-ears-run-nypd-s-use-sound-cannons-challenged-n1008916. Last accessed on January 1st 2021.

[Khamparia2019] Khamparia, A., and Gupta, D. and Nguyen, N.G. and Khanna, A. and Pandey, B. and Tiwari, P. “Sound Classification Using Convolutional Neural Network and Tensor Deep Stacking Network”, IEEE International Conference on Network Infrastructure and Digital Content (IC-NIDC), pp. 7717-7727, 2019.

[Khan2019] Khan, A. and Sohail, A. and Zahoora, U. and Qureshi, A. A survey of the recent architectures of deep convolutional neural networks. arXiv preprint arXiv:1901.06032, 2019.

[Kiapuchinski2012] Kiapuchinski, L. and Kaestner. Spectral Noise Gate Technique Applied to Birdsong Preprocessing on Embedded Unit. IEEE International Symposium on Multimedia, pp. 24-27, 2012.

[Kim2020] Kim, Y. and Han, D. and Kim, C. and Yoo, H. “”A 0.22-0.89 mW Low-Power and Highly-Secure Always-On Face Recognition Processor with Adversarial Attack Prevention”, IEEE Transactions on Circuits and Systems II, Volume 67, Issue 5, 2020.

[Koerich2019] Koerich, K. and Esmailpour, M. and Abdoli, S. and Britto Jr., A. “Cross-Representation Transferability of Adversarial Attacks: From Spectrograms to Audio Waveforms”, IEEE International Joint Conference on Neural Networks, 2020.

[Kwon2019] Kwon, H. and Kim, Y. and Yoon, H. and Choi, D. Selective audio adversarial example in evasion attack on speech recognition system. IEEE Transactions on Information Forensics and Security, pp. 526-538, 2019.

[Kubo2019] Kubo, Y. and Trappenberg, T. Mitigating overfitting using regularization to defend networks against adversarial examples. Canadian Conference on Artificial Intelligence, pp. 400-405, 2019.

[Kumar2020] Kumar, K. and Vishnu, C. and Mitra, R. and Mohan, C. “Black-box Adversarial Attacks in Autonomous Vehicle Technology”, IEEE Applied Imagery Pattern Recognition Workshop, 2020.

[Kurakin2016] Kurakin, A. and Goodfellow, I. and Bengio, S. Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533, 2016.

[Kurakin2017] Kurakin, A. and Goodfellow, I. and Bengio, S, “Adversarial Machine Learning at Scale”, ICLR, 2017.

[Larrucea2017] Larrucea, X., and Combelles, A., and Favaro, J., and Taneja, K. “Software Engineering for the Internet of Things”. IEEE Software, pp. 24-28, 2017.

[Larson2011] Larson, E. and Lee, T. and Liu, S. and Rosenfeld, M. and Patel, S. “Accurate and privacy preserving cough sensing using a low-cost microphone”. Proceedings of the 13th international conference on Ubiquitous computing, pp. 375-384, 2011.

[Li2016] Li, Y. and Liu, G. “Sound classification based on spectrogram for surveillance applications”, IEEE International Conference on Network Infrastructure and Digital Content (IC-NIDC), pp. 293-297.

[Li2020] Li, Z. and Wu, Y. and Liu, J. and Chen, Y. and Yuan, B. “AdvPulse: Universal, Synchronization-Free, and Targeted Audio Adversarial Attacks via Subsecond Perturbations”, ACM SIGSAC Conference on Computer and Communications Security, pp. 1121-1134.

[Li2021] Li, Y. and Wang, Y. and Liu, X. and Shi, Y. and Shih, S. Enabling Real-time On-chip Audio Super Resolution for Bone Conduction Microphones. arXiv preprint arXiv:2112.13156, 2021.

[Lijima2019] Lijima, R. and Minami, S. and Zhou, Y. and Takehisa, T. “Audio Hotspot Attack: An Attack on Voice Assistance Systems Using Directional Sound Beams and Its Feasibility”, IEEE Transactions on Emerging Topics in Computing, 2019.

[Lim2017] Lim, H. and Park, J. and Han, Y. Rare sound event detection using 1D convolutional recurrent neural networks. Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop, pp. 80-84, 2017.

[Liu2019] Liu, S. and Wu, H. and Lee, H. and Meng, H. “Adversarial Attacks on Spoofing Countermeasures of Automatic Speaker Verification”, IEEE Automatic Speech Recognition and Understanding Workshop, 2019.

[Madry2018] Madry, A. and Makelov, A. and Schmidt, L. and Tsipras, D. “Towards Deep Learning Models Resistant to Adversarial Attacks”, ICLR, 2018.

[Mao2020] Mao, J. and Zhu, S. and Xuan, D. and Lin, Q. and Liu, J. Watchdog: Detecting Ultrasonic-based Inaudible Voice Attacks to Smart Home Systems. IEEE Internet of Things Journal, 2020.

[Matos2006] Matos, S. and Birring, S. and Pavord, I. and Evans, H. Detection of cough signals in continuous audio recordings using hidden Markov models, IEEE Transactions on Biomedical Engineering, pp. 1078-1083, 2006.

[Matsuoka2020] Matsuoka, Y. and Nongpiur, R. and Dixon, M. “Method and system for detecting an audio event for smart home devices”, Google Patents, 2020.

[Mendes2020] Mendes, E. and Hogan, K. “Defending Against Imperceptible Audio Adversarial Examples Using Proportional Additive Gaussian Noise”, 2020.

[MIMII2019] Purohit, H. and Tanabe, R. and Ichige, K. and Endo, T. and N, Y. and Suefusa, K. and Kawaguchi, Y. Sound Dataset for Malfunctioning Industrial Machine Investigation and Inspection. Available at https://zenodo.org/record/3384388#.YCf3XmhKhjE. Last accessed on May 2021.

[Miraz2015] Miraz, M. and Ali, M. and Excell, P. and Picking, R. “A review on Internet of Things, Internet of Everything and Internet of Nano Things”. Internet Technologies and Applications, 2015.

[Mizokami2010] Popular Mechanics. So What Is This Secretive Chinese Sonic Weapon Exactly?. Available at www.popularmechanics.com/military/. Last Accessed in January 1st 2021.

[Montillet2013] Montillet, J. and Tregoning, P. and McClusky, S. and Yu, K. Extracting White Noise Statistics in GPS Coordinate Time Series. IEEE Geosci. Remote. Sens. Letters, pp. 563-567, 2013.

[Moosavi2016] Moosavi-Dezfooli, S. and Fawzi, A. and Frossard, P. “Deep-Fool a Simple and Accurate Method to Fool Deep Neural Networks”, Conference on Visual Processing, 2016.

[Nest2021a] Nest. Nest Hub. Available at www.store.google.com/us/product/google_nest_hubl. Last accessed on January 1st 2021.

[Nest2021b] Nest. Nest Mini. Available at www.store.google.com/product/google_nest_mini. Last accessed on January 1st 2021.

[Nwankpa2015] Nwankpa, C. and Ijomah, W. and Gachagan, A. and Marshall, S. Activation Functions: Comparison of Trends in Practice and Research for Deep Learning. Proceedings of the Machine Learning: Computer Vision and Pattern Recognition, 7-12, 2015.

[Nytimes2018] Mervosh, S. Nearly 40,000 People Died From Guns in U.S. Last Year, Highest in 50 Years. Available at https://www.nytimes.com/2018/12/18/us/gun-deaths.html. Last accessed on August 2020.

[Oh2019] Oh, S. et al. “An Acoustic Signal Processing Chip With 142-nW Voice Activity Detection Using Mixer-Based Sequential Frequency Scanning and Neural Network Classification”, IEEE Journal of Solid-State Circuits, vol. 54, no.11, pp. 3005-3016, 2019.

[Olivier2021] Olivier, R. and Raj, B. and Shah, M. “High-Frequency Adversarial Defense for Speech and Audio, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021.

[Page1995] Page, E. and Sharkey, B. “SECURES: System for Reporting Gunshots in Urban Environments”, Symposium on OE Aerospace Sensing and Dual Use Photonics, 1995.

[Pal2021] Pal, M. and Jati, A. and Peri, R. “Adversarial Defense for Deep Speaker Recognition using Hybrid Adversarial Training”, IEEE International Conference on Acoustics, Speech and Signal Processing, 2021.

[Peng2009] Peng, Y. and Lin, C. and Sun, M. and Tsai, K. “Healthcare audio event classification using hidden Markov models and hierarchical hidden Markov models”. IEEE International conference on multimedia and expo, pp. 1218-1221, 2009.

[Perez2017] Perez, L. and Wang, J. The Effectiveness of Data Augmentation in Image Classification using Deep Learning. arXiv:1712.04621v1, 2017.

[Petridis2010] Petridis, S. and Giannakopoulos, T. and Perantonis, S. “A multi-class method for detecting audio events in news broadcasts”. Hellenic Conference on Artificial Intelligence, pp. 399-404, 2010.

[Pillos2016] Pillos, A. and Alghamidi, K. and Alzamel, N. and Pavlov, V. and Machanavajhala, S. “A real-time environmental sound recognition system for the Android OS”, Proceedings of Detection and Classification of Acoustic Scenes and Events, 2016.

[Pocketlint2021] Pocket Lint. What is Google Home, what can it to and how does it work. Available at www.pocket-lint.com/smart-home/news/google/137665-what-is-google-home-what-can-it-do-and-how-does-it-work. Last accessed on January 1st 2021.

[Policyadvice2021] Policy Advice. 20+ Exciting Google Home Statistics to Prepare You for 2021. Available at https://analyticsindiamag.com/how-machine-learning-rocked-googles-hardware-event-this-year. Last accessed on December 2019.

[Prasadh2017] Prasadh, K. and Natrajan, S. and Kalaivani, S. “Efficiency analysis of noise reduction algorithms: Analysis of the best algorithm of noise reduction from a set of algorithms”. International Conference on Inventive Computing and Informatics, pp. 1137-1140, 2017.

[Rabaoui2008] Rabaoui, A. and Kadri, H. and Ellouze, N. “New approaches based on One-Class SVMS for impulsive sounds recognition tasks”, IEEE Workshop on Machine Learning for Signal Processing, pp. 285-290, 2008.

[Rajaratnam2018] Rajaratnam, K. and Kalita, J. “Noise flooding for detecting audio adversarial examples against automatic speech recognition”, IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), pp. 197-201, 2018.

[Reddy2018] Reddy, R. and Mamatha, Ch. and Reddy, R.G. A Review on Machine Learning Trends, Application and Challenges in Internet of Things. International Conference on Advances in Computing, Communication and Informatics, 2018.

[Ring2021] Ring. “Always Home Cam A New Kind of Camera that Flies”. Available at https://ring.com/always-home-cam-flying-camera. Last accessed on October 2021.

[Robinson2018] Robinson, M. and Gould, S. and Lee, S. There have been 307 mass shootings in the US so far in 2018 — here's the full list. Available at https://www.businessinsider.com/how-many-mass-shootings-in-america-this-year-2018-2. Last accessed in April 30, 2019.

[Roelandts2013] Roelandts, Tom. What Is a Spectrogram. Available at www.tomroelandts.com/articles/what-is-a-spectrogram. Last accessed in January 1st 2021.

[Roy2018] Roy, N. and Shen, S. and Hassanieh, H. and Choudhury, R. Inaudible voice commands: The long-range attack and defense. Symposium on Networked Systems Design and Implementation, pp. 301-305, 2020.

[Sainburg2018] Sainburg, T. Noise reduction using spectral gating in python. Available at https://timsainburg.com/noise-reduction-python.html. Last accessed on May 2021.

[Sallo2020] Sallo, R. and Esmaeilpour, M. and Cardinal, P. “Adversarially Training for Audio Classifiers”, International Conference on Pattern Recognition, 2020.

[Schonher2018] Schnherr, L. and Kohls, K. and Zeiler, S. and Holz, T. and Kolossa, D. “Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding”, arXiv preprint arXiv:1808.0566, 2018.

[SeeSound2021] Wavio. See Sound. Available at: www.see-sound.com/devices/. Last accessed on January 1st, 2021).

[Shailaja2018] Shailaja, K. and Seetharamulu, B. and Jabbar, M. Machine Learning in Healthcare: A Review. Second International Conference on Electronics, Communication and Aerospace Technology, 2018.

[Shrestha2014] Shrestha, N. and Kubler, S. and Framling, K. “Standardized Framework for Integrating Domain-Specific Applications into the IoT”. International Conference on Future Internet of Things and Cloud, 2014.

[Shiekh2017] Shiekh, A. and Tahir, M. and Uppal, M. “Accurate gunshot detection in urban environments using blind deconvolution”, International Multi-topic Conference (INMIC), 2017.

[Shijie2017] Shijie, J. and Ping, W. and Peiyi, J. and Siping, H. Research on data augmentation for image classification based on convolution neural networks. Chinese Automation Congress (CAC), 2018.

[Shooter2020] Shooter. “Shooter Detection Systems”.
Available at https://shooterdetectionsystems.com. Last accessed on May 2021.

[Showen1997] Showen, R. “Surveillance and Assessment Technologies for Law Enforcement}”. International Society for Optics and Photonics, pp. 130-139, 1997.

[Singh2014] Singh, D. and Tripathi, G. and Jara, A. A survey of Internet-of-Things: Future vision, architecture, challenges and services. IEEE World Forum on Internet of Things, 2014.

[Somvanshi2016] Somvanshi, M. and Chavan, P. “A review of machine learning techniques using decision tree and support vector machine”. International Conference on Computing Communication Control and Automation, 2016.

[Song2018] Song, C. And Cheng, H. And Li, S. And Wu, C. And Wu, Q., Chen, Y., Li, H. MAT: A Multi-strength Adversarial Training Method to Mitigate Adversarial Attacks. IEEE Computer Society Annual Symposium on VLSI (ISVLSI), pp. 476-481, 2018.

[Soundmachines2021] The Strategist. The Best Sound Machines on Amazon, According to Hyperenthusiastic Reviewers. Available at www.nymag.com/strategist/article/best-sound-machines-noise-machines.html. Last accessed in January 1st 2021.

[Srivastava2014] Srivastava, N. and Hinton, G. and Krizhevsky, A. and Sutskever, I. and Salakhutdinov, R. Dropout A Simple Way to Prevent Neural Networks from Overfitting. Journal of Mach. Learn. Res., pp. 1919-1958, 2014.

[Su2019] Su, J. and Vargas, D. and Sakurai, K. One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation, pp. 828-841, 2019.

[Subramanian2020] Subramanian, V. and Pankajakshan, A. and Benetos, E. and Xu, N. and McDonald, S. and Sandler, M. A study on the transferability of adversarial attacks in sound event classification. IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 301-305, 2020.

[Szegedy2013] Szegedy, C. and Zaremba, W. and Sutskever, I. “Intriguing Properties of Neural Networks”, arXiv preprint, 2013.

[Tangkawanit2018] Tangkawanit, S. and Pinthong, C. and Kanprachar, S. “Development of gunfire sound classification system with a smartphone using ANN”, International Conference on Digital Arts, Media and Technology (ICDAMT), pp. 168-172, 2018.

[Thakkar2018] Thakkar, V. and Tewary, S. and Chakraborty, C. Batch Normalization in Convolutional Neural Networks—A comparative study with CIFAR-10 data. Proceedings of the Fifth International Conference on Emerging Applications of Information Technology, pp. 12-13, 2018.

[UrbanSound2014] Salamon, J. and Jacoby, C. A Dataset and Taxonomy for Urban Sound Research. Available at www.justinsalamon.com. Last accessed in May 2020.

[Yamashita2018] Yamashita, R. and Nishio, M. and Do, R. and Togashi, K. “Convolutional neural networks: An overview and application in radiology”. Insights Imaging, pp. 611-629, 2018.

[Yan2020] Yan, Q. and Liu, K. and Zhou, Q. and Guo, H. and Zhang, N. “Surfing Attack: Interactive Hidden Attack on Voice Assistants Using Ultrasonic Guided Waves”, Network and Distributed Systems Security, 2020.

[Vafeiadis2020] Vafeiadis, A. and Votis, K. and Giakoumis, D. and Tzovaras, D. and Chen, L. and Hamzaoui, R. “Audio content analysis for unobtrusive event detection in smart homes”, Engineering Application of Artificial Intelligence, pp. 103226, 2020.

[Vashuki2012] Vasuki, P. and Bhavana, C. and Mohamed, S. and Lakshmi, E. Automatic noise identification in images using moments and neural network. International Conference on Machine Vision and Image Processing (MVIP), pp. 14-15, 2012.

[Venkatesh2017] Venkatesh, J. and Aksanli, B. and Chan, C. and Akyurek, A. and Rosing, T. Scalable-Application Design for the IoT. IEEE Software, pp 62-70, 2017.

[Wang2019] Wang, J. And Zhang, H. Bilateral Adversarial Training: Towards Fast Training of More Robust Models Against Adversarial Attacks.
IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6628-6637, 2019.

[Wei2005] Wei, L. and Yang, Y. and Nishikawa, R.M. and Jiang, Y. A study on several Machine-learning methods for classification of Malignant and benign clustered microcalcifications. IEEE Transactions on Medical Imaging, 2005.

[Zapsplat2021] McKinney, A. and Harris, G. Free sound effects & royalty free music. Available at http://https://www.zapsplat.com. Last accessed on May 2021.

[Zhang2017] Zhang, G. and Yan, C. and Ji, X. and Zhang, T. and Zhang, T. and Xu, W. Dolphinattack: Inaudible voice commands. Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, pp. 103-117, 2017.

[Zhang2021] Zhang, W. and Zhao, S. and Li, J. and Cheng, X. “Attack on Practical Speaker Verification System using Universal Adversarial Perturbations”, IEEE International Conference on Acoustics, Speech and Signal Processing, 2021.

[Zhou2017] Zhou, H. and Song, Y. and Shu, H. Using deep convolutional neural network to classify urban sounds.

TENCON 2017 - 2017 IEEE Region 10 Conference, pp. 3089-3092, 2017.