Defending Adversarial Attacks on Cloud-aided Automatic Speech Recognition Systems

Jiajie Zhang
j.zhang41@lancaster.ac.uk
Lancaster University

Bingsheng Zhang
b.zhang2@lancaster.ac.uk
Lancaster University

Bingcheng Zhang
zhangbincheng1997@gmail.com
Jinan University

ABSTRACT

With the advancement of deep learning based speech recognition technology, an increasing number of cloud-aided automatic voice assistant applications, such as Google Home, Amazon Echo, and cloud AI services, such as IBM Watson, are emerging in our daily life. In a typical usage scenario, after keyword activation, the user's voice will be recorded and submitted to the cloud for automatic speech recognition (ASR) and then further action(s) might be triggered depending on the user's command(s). However, recent researches show that the deep learning based systems could be easily attacked by adversarial examples. Subsequently, the ASR systems are found being vulnerable to audio adversarial examples. Unfortunately, very few works about defending audio adversarial attack are known in the literature. Constructing a generic and robust defense mechanism to resolve this issue remains an open problem. In this work, we propose several proactive defense mechanisms against targeted audio adversarial examples in the ASR systems via code modulation and audio compression. We then show the effectiveness of the proposed strategies through extensive evaluation on natural dataset.

CCS CONCEPTS

• **Security and privacy → Security services**; *Systems security*; • **Computing methodologies → Speech recognition**.

KEYWORDS

Adversarial Examples; Deep Learning; Cloud-aided Speech Recognition

ACM Reference Format:
Jiajie Zhang, Bingsheng Zhang, and Bingcheng Zhang. 2019. Defending Adversarial Attacks on Cloud-aided Automatic Speech Recognition Systems. In *7th Int'l Workshop on Security in Cloud Computing (SCC '19), July 8, 2019, Auckland, New Zealand*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/XXXXXX.XXXXXX

1 INTRODUCTION

Automatic Speech Recognition (ASR) technology aims at converting the phrases or words spoken by human into text. During the past years, cloud-aided ASR technology has been successfully commercialized and gradually moving from the laboratory to our daily life, especially in the areas of smart home. The related applications, such as Apple's Siri, Google Home, and Amazon Echo, have been adopted to millions of end users. In a typical usage scenario, after keyword activation, the user's voice will be recorded and submitted to the cloud for speech recognition and then further action(s) might be triggered depending on the user's command(s).

Meanwhile, an increasing number of cloud AI outsourcing services are emerging to facilitate computationally limited devices. Backed by Google machine learning techniques and Google Cloud Platform, Google Cloud Speech-to-Text [13] and Dialogflow [15] provide API to developers, and enable real-time audio processing. Another Speech to Text system from Access Watson services on IBM Cloud has been used in various applications, such as customer self-service virtual assistant to speed-up customer response. Moreover, Speechmatics [46] offers cloud-based service for multiple language ASR even in a noisy environment.

In common practice, to solve the problem of the cumbersome multistep exercise including hand-engineered processing when building ASR systems, *end-to-end* learning is adopted to simplify the sophisticated pipelines and supersede these processing stages by using deep learning technologies [1, 43]. However, recent researches show that these deep learning-driven systems are vulnerable to adversarial attacks [49]. In those attacks, the attackers maliciously inject tailor-made small perturbations into the source data, which cannot be detected by human recognition, while they are able to compromise the integrity of the decisions made by the deep learning systems or algorithms.

In the literature, the concerns of adversarial examples have been raised not only in a broad range of image processing tasks, such as image classification [4, 20, 28, 38, 48], semantic segmentation [3, 19], human pose estimation [11], and object detection [52], but also in reinforcement learning agent [5]. Moreover, adversarial attacks are also emerged in the audio processing domain recently [2, 10, 12]. More specifically, the malicious attackers can construct targeted and/or untargeted audio adversarial examples to launch adversarial attacks against ASR systems, which draws a widely public attention to the security problems caused by these types of attacks.

These kinds of attacks can be classified by the goals and phases. In terms of the attacking goals, the malicious attackers can cause the deep learning algorithms to make wrong decisions based on their wishes, that's the so-called targeted attacks, whereas, the untargeted attacks mean that the decisions can be anything but the normal one. As for the attacking phase, there are two broad types including the evasion attacks [20, 49] and data poisoning attacks [32, 33, 51]. Evasion attacks are commenced during the deploying phase, where attackers query the algorithms by adding crafted

SCC '19, July 8, 2019, Auckland, New Zealand

ACM ISBN 978-1-4503-6788-2/19/07…$15.00
https://doi.org/10.1145/XXXXXX.XXXXXX

noise in the input data. A classical evasion attack [18, 38] is that the attacker alters the STOP sign, which is still a STOP sign from human perceptional perspective, but recognised as a SPEED sign by the recognition system of autonomous vehicle, this may lead to serious traffic accidents and even life-threatening matters. Another kind is data poisoning attacks which mainly happen during the training phase, since it's obvious that the real world interacted with the deep learning algorithms is continuously changing, the applications based on deep learning algorithms should be never-ending learning and continual [35] or life-long learning [29, 45]. The attackers can take advantages of this situation by injecting wrong data in the training data to make the algorithms do error modeling. As an example, Tay bot developed by Microsoft learned from the people during interaction and made offensive or racist statements [40].

Most known defense strategies against adversarial examples mainly focus on image processing, however, those types of strategies cannot be directly applied to audio processing. The main difference between audio and image processing is rate of statistical redundancies. Audio data shows much more repetitions than images, which makes techniques non-portable from one context to another. As a practical example, consider a 16-bit coded WAC audio for example, the range of the sample values is $[-32768, 32767]$. Comparing with the pixel value range of the image $[0, 255]$, the former one is 256 times wider than the latter one, which significantly increases the difficulty of detecting the adversarial attacks. Unfortunately, with regard to the defense strategies against audio adversarial examples in cloud-aided ASR systems, very few progress has been made. We here ask the following challenging question:

> *Is it possible to introduce a proactive defense mechanism that can be integrated with the off-the-shelf cloud-aided ASR systems to protect them from being abused by the potential audio adversarial examples?*

We very much expect an affirmative answer because we need to ensure that the robustness of those cloud-aided ASR systems to enable their wide adoption in security critical applications.

Our contributions. In this work, we, for the very first time, study the proactive defense strategies for the cloud-aided ASR systems against audio adversarial examples. The defense strategy is independent of the protected cloud-aided ASR systems, which means that our proposed defense mechanism can be readily deployed to protect extensive off-the-shelf audio processing systems. Furthermore, the structure of the protected cloud-aided ASR systems needn't to be modified. We summarize our main contributions as follows:

- We initiate the study of proactive defense strategy to mitigate the risk of adversarial examples in cloud-aided ASR systems. In particular, we propose the world's first proactive defense mechanism to protect existing cloud-aided ASR systems from adversarial attacks. The proposed protection mechanism can be deployed on existing systems and run as an online watchdog to detect and defend with little computational overhead.
- This work also studies and formulates the properties of audio adversarial examples and security model of a defense strategy, enabling us to directly measure the effectiveness and efficiency of the defense approaches.
- With evaluated by a natural dataset, the effectiveness of our proposed defense approach is measured by the recognition time, the edit distance of the transcribed text and the difference of the processing audio. To the best of our knowledge, this is the first study aiming at mitigating the effects of adversarial attacks on cloud-aided ASR systems.

Organisation. The rest of this article is organized as follows. Problem statement of audio adversarial examples, security model, and adversarial attack are presented in Section 2. Next, the proposed active transformation defense approach is described in Section 3, followed by the experimental analysis based on a real-world radio dataset. In Section 5, we review the related work about generating and defending adversarial examples. Finally, we draw our conclusion and discuss future research directions.

2 PROBLEM STATEMENT

2.1 Audio Adversarial Examples

Before giving the definitions about generating audio adversarial examples against ASR systems as described in [10], we will make a succinct introduction of the recognition process of the ASR systems [23], which helps the readers understand the attack process and our defense strategy. Considering $\mathbb{R}$ is the set of all the audio samples in the input domain space, X is a single frame of the input audio in $\mathbb{R}$, $\mathbb{Y}$ is the range of the recognition results, such as the a to z characters, the space and the special ϵ of the output domain $\mathbb{O}$. A normal neural network in ASR systems is to return a probability distribution over $\mathbb{O}$, based on the input sequences, which can be insulted as a function $f : \mathbb{R} \rightarrow [0, 1]^{|\mathbb{Y}|}$. Relatively, $f(\cdot)$ means a probability distribution over the characters of every frame. To get the probability distribution of all the phrases, Connectionist temporal classification (CTC) [21] is applied to train the ASR systems without knowing the alignment between the input and output sequences, by minimizing the CTC loss:

$$\text{CTC-Loss}(f(x), p) = -\log \Pr[p|f(x)]$$

where $\Pr(p|f(x))$ stands for the probability of a given phrase p over the distribution $y = f(x)$.

In terms of the attacking procedure, following the work in [10], targeted adversarial examples are considered in this article. Firstly, we define $\mathbb{N} = x|x \in \mathbb{R}$, in this definition, x occurs naturally with regard to the ASR systems. We limit the scope of x to study the manifold of the input domain, as stated in [26]. Considering the space of $\mathbb{R}$ is much larger than $\mathbb{N}$, we can formalize $\mathbb{R} \rightarrow \mathbb{N} \cup \perp$, where $\perp$ represents that the input of x is unlikely to be judged from the data generation process of $\mathbb{R}$. The manifold of input space is one of the issues that needs clarifying, the challenge is to determine whether there are differences between the natural audio examples and the audio adversarial examples. Additionally, we follow Carlini and Wagner's work, choosing the Mozilla Common Voice dataset as the natural dataset for audio recognition.

An audio adversarial example x' have three definitions in our discussion:

(1) The perturbation of x' and x should be imperceptible to human, which can be quantified using the distortion in Decibels (dB), as the opposite idea against the pointless adversarial fooling examples in [6, 37].
(2) x' must be assigned by the neural network with a specific label chosen by the adversary, not the type of untargeted adversarial examples. As stated in [10], the untargeted audio adversarial examples are not usually so interesting and do little help to the adversary.
(3) x' should follow these two constraints: $f(x') \neq f(x)$ and $x \in \mathbb{R} \setminus \mathbb{N}$, $f(x') \neq f(x)$, which point out that when ASR systems make a wrong decision given a nature input sample, this source of error is not considered as the so called adversarial example, since no system is perfect. If this is the case, then the attacker needs to find out all the natural examples which can lead to system error through brute force search. That will be a very time-consuming and laborious collection task, which requires the annotation of all natural samples by human. Therefore, we have introduced the constrain of $f(x') \neq f(x)$ to limit the countermeasures causing errors. Only artificial modifications can be used to deceive the system rather than a natural sample.

2.2 Security Model

We assume that the attacker knows everything of the ASR systems that he wishes to attack, which can be called white box attack. The knowledge of the attacker contains the model structure, parameters, and the training procedure, excluding the defense approach d_f. In this type of attack, the attackers are allowed to have the maximum power, which is considered as the most taken situation in prior works [20, 25, 38]. The adversary transcripts the given audio waveform x to x', by constructing $x' = x + \sigma$, additionally, x' sounds similar to x but $f(x') = y$.

What's more, the defense strategy in our secure system knows nothing about how the adversary generates the audio adversarial examples. More specifically, we give a system security definition as follows: A defense strategy against the adversarial attack is to construct a filter which destructs the robustness of audio adversarial examples. We define the original radio waveform is a random variable c with probability distribution of P_c, the attacking process is a function defined over c, and P_a stands for the probability distribution of all perturbations produced during the attack. Besides, we formalize the defense process as $P_a \rightarrow P_f$, by generating the probability distribution P_f of all the adversarial examples after defense reconstructing. This defense strategy is to make the reconstruction probability distribution P_f similar to the original probability distribution P_c.

2.3 Adversarial Attack

In this section, we briefly introduce the adversarial attack presented in [10]. At first, the attacking process is to solve the formulation:

$$\begin{aligned} \textbf{minimize} \quad & \mathrm{dB}_x(\delta) \\ \textbf{subject to} \quad & f(x+\delta) = t \\ & x + \delta \in [-M, M] \end{aligned}$$

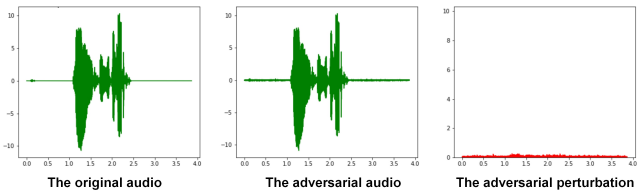

Figure 1: Adversarial audio examples and corresponding perturbation.

where M is the maximum representable value (2^{15} used in this article). By clipping the value of δ, the constraint can be handled. To solve non-linearity difficulty of $f(x+\delta) = t$, the non-trivial loss function constructing, and the l_∞ distortion metric, the authors applied the trick of *C&W* attack, and initially solve the formulation:

$$\begin{aligned} \textbf{minimize} \quad & |\delta|_2^2 + c \cdot l(x+\delta, t) \\ \textbf{subject to} \quad & \mathrm{dB}_x(\delta) \leq \tau \\ & x + \delta \in [-M, M] \end{aligned}$$

where c is a hyperparameter that balances the length and deceiving effect, the bound δ is defined to have distortion at most τ, and τ is initialized to some sufficiently large constant.

The loss function is the CTC loss:

$$l(x+\delta, t) = \text{CTC-Loss}(x+\delta, t)$$

By reducing τ and resuming minimization until no solution δ^* can be found, the optimization results with the minimum distortion are returned, with satisfying the box constraints to be a valid audio.

Normally, for a fixed audio waveform x, the adversarial attack is to find the minimum perturbation δ that is small enough in length and can deceive the ASR systems at the same time.

3 COUNTERMEASURE DESIGN

The adversarial attacks may change the specific statistics of the input audio waveform to fool the ASR systems, in fact, the adversarial perturbation has a specific structure, as shown in Fig.1.

We design a proactive defense mechanism to recover the structure of the original input and remove the perturbations, then investigate whether these changes can eliminate the impact of the adversarial attacks. As shown in Fig.2, we construct a filter and a detector. The audio, as an input will be processed by two components simultaneously. The first one is the ASR system, the output recognition result will be sent to the detector. The second part is the filter that will process the audio firstly and send the audio to ASR system. After that, the detector gets two types of recognition results, it will calculate the difference between these two recognition results. Once the difference is larger than the predefined threshold, the detector will alarm and classify the input audio as an adversarial example. If the difference is below the threshold, the input sample will be marked as a clean one, then the output text recognised by ASR systems will be sent to the user.

The detector catches the suspicious input examples by measuring the difference of the original audio waveform and the processed one. Once the difference is higher than the threshold chosen by

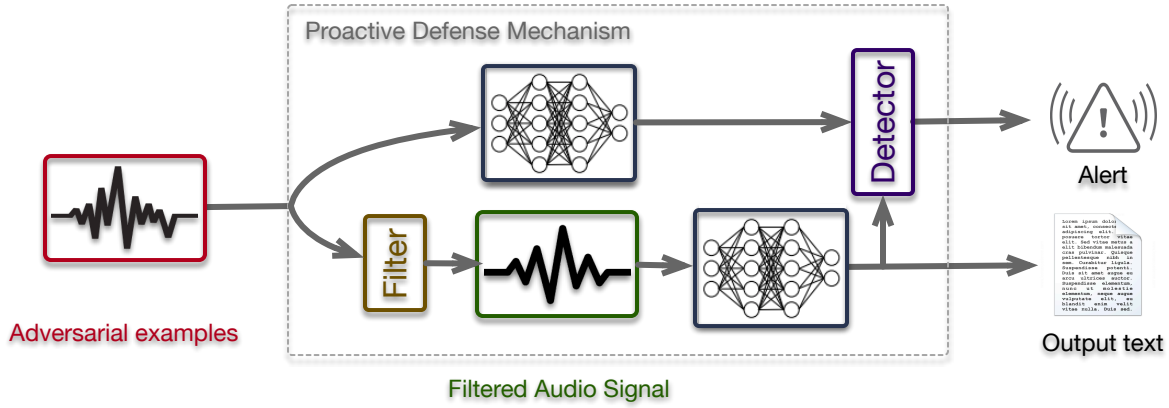

Figure 2: The proposed proactive defense mechanism.

the user of the defense system, filter is activated to mitigate the adversarial attack. After filtering the adversarial audios, they are directly recognised by the server-aided ASR system. In terms of filter, we propose two categories of transformation technologies as a pre-processing step before running the server-aided ASR system: (i). Code Modulation, (ii). Audio Compression.

3.1 Code Modulation

Code Modulation is divided into two parts in our work. Firstly, we remove the heading of the input radio waveform, and apply the G.729 [42] narrow-band vocoder-based audio data compression algorithm to compress the processed radio into bits. Secondly, pulse-code modulation (PCM) [16] is used to reconstruct the audio waveform based on the bits with adding the aforementioned heading.

G.729 is an audio coding algorithm defined by ITU Telecommunication Standardization Sector (ITU-T), with a frame length of 10 milliseconds and 25 ms point-to-point delay. G.729 is based on the Code-Excited Linear Prediction (CELP) mode and applies CS-ACELP (Conjugate-Structure Algebraic Code Excited Linear Prediction) to encode the audio at a baud rate of 8 Kbps, which is almost always used for Voice over IP (VoIP) with low bandwidth requirements.

PCM is a method for digitally representing sampled analog signals. It is a standard form of digital audio in computers, compact discs, digital phones, and other digital audio applications. In the PCM stream, the amplitude of the analog signal is sampled at regular intervals. Each sample is quantized to the nearest value within the digital step range, and expresses a good reproduction of the perceptual quality of the original uncompressed audio. On this account, the amount of data required to represent the audio signal recorded as PCM is tremendously reduced.

3.2 Audio Compression

To provide resilience to attack in the study of image adversarial examples, the authors in [14, 17, 22] explored systematic JPEG compression as a pre-processing step. Their method significantly reduces adversarial perturbation. For the reason that adversarial attacks are deployed by introducing adversarial perturbations which are beyond human psychovisual awareness, we apply audio compression as one of the active transformation approaches to remove the artifacts in the adversarial examples.

As for audio compression, MP3 has been one of the most popular and trusted audio compression standards. A core principle behind MP3 compression is the lossy data compression, which encodes data using inaccurate approximations and partially discarded data. This type of technology can greatly reduce the file size compared to uncompressed audio. MP3 compression records the residual audio information in a space-saving manner by reducing (or approaching) the accuracy of certain sound components considered to exceed the hearing capabilities of human. This approach is commonly referred to as perceptual coding or psychoacoustic modeling, which reduces the storage space without perceptible difference.

4 IMPLEMENTATION AND EVALUATION

In this section, we perform four experiments to evaluate the efficiency and the properties of the active transformation defense strategy described in Section 3, against the audio adversarial attacks stated in [10] with a natural dataset: the Common Voice [36]. As described in our security model, the adversary is considered to access the model architecture of the ASR systems and parameters, but unknown to the specific defense strategy.

4.1 Experimental Setup

In our experiment, we choose the cv-invalid-dev sub-dataset of the Common Voice as the natural audios, which contains 4076 samples. We firstly adopt the adversarial attacks methods released by Carlini to generate the audio adversarial examples based on CTC loss, followed by the experimental setting in [10]. It should be emphasized that, when setting the length of the adversarial target, we should consider about the Mel-Frequecny Cepstrum (MFC) used in the processing of reducing the input dimensionality. Because the radio waveform are split into 50 frames per second, which relatively limits the maximum density of a audio waveform at 50 characters per second. In this work, the target translation sentence

Table 1: The precision measurements of proposed defense approach in different thresholds setting.

Defense Approach	Threshold	FNR	FPR
Code Modulation	0.4375	0.0462	0.0002
MP3 Compression 64Kb/s	0.3556	0.1069	0.0002
MP3 Compression 96Kb/s	0.4219	0.0494	0.0002
MP3 Compression 128Kb/s	0.4219	0.0509	0.0002

(the recognition results of generated audio adversarial examples) is fixed with 64 bytes during the whole experiment, and finally we get 4070 adversarial radio waveforms (discard 6 failed audio adversarial examples).

According to the aforementioned security model defined in Section 3, the security definition of the protected system is related to the reconstruction probability distribution P_f and the original probability distribution P_c, we measure the correlation in terms of Minkowski Distance, Signal-Noise Ratio (SNR), and edit distance.

The Minkowski distance is a metric in the norm vector space which generalizes a wide range of distances such as the Hamming distance, the Euclidean distance, the geometric distance, and the normalized harmonic distance. To measure the Minkowski distance of dimension n, we construct a mapping d_m: $\mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R}$, such that:

$$d_m(A,B) = \left(\frac{1}{n}\sum_{i=1}^{n}|a_i - b_i|^{\lambda}\right)^{1/\lambda} \tag{1}$$

where a_i and b_i are the ith elements of the sets A and B, λ is a variable parameter such that $\lambda \in (-\infty, +\infty)$. In particular, if $\lambda = 1$, the Minkowski distance is precisely the Hamming distance; $\lambda = 2$ stands for the Euclidean distance; when $\lambda \to \infty$, the Minkowski distance can be considered as Chebyshev distance. We mainly consider the Chebyshev distance measurement in our experiment to evaluate the distance among the original audios, the adversarial audios and the transformed audios after filtrating relatively.

SNR is often used in science and engineering to compare the level of desired signal and the background noise, which stands for the ratio of signal power to noise power. Normally, when the ratio is greater than 1:1 (or higher than 0 dB), it means that there are more signals than noise. In our work, we use the difference between two audios to verify the performance of proposed machnism. SNR is defined as:

$$\mathrm{SNR(dB)} = 10\log_{10}\left(\frac{P_{\mathrm{signal}}}{P_{\mathrm{adv}}}\right) = 10\log_{10}\left(\frac{A^2_{\mathrm{signal}}}{A^2_{\mathrm{adv}}}\right) \tag{2}$$

where P_{signal} is the power of signal, P_{adv} is the power of adversarial perturbation, A_{signal} is the amplitude of signal, and A_{adv} is the amplitude of adversarial perturbation.

In the standard approach to ASR systems, given the acoustic information A, the optimization goal is to find the sentence hypothesis that maximizes the posterior probability $P(W|A)$ of the word sequence W. However, the general difficulty of measuring performance is that the recognized word sequence may be of different length than the reference word sequence (presumably correct). For this reason, the distance metrics like Euclidean distance are no longer applicable to quantify the mutual relationship of two word sequences. The commonly used performance metric in ASR systems is edit distances $WE(W,R)$ between a hypothesis W and the reference string R. $WE(W,R)$ is defined as the number of substitutions, deletions, and insertions relative to R in the alignment of two strings, which minimizes the weighted combination of these types of errors. In this article, to more fully measure our defensive performance, we choose Levenshtein distance and the Word error rate (WER) which is derived from the Levenshtein distance as the measurements.

Table 2: SNR and Chebyshev distance evaluation.

Defense Approach	Metric	adv_advcode	orig_adv	orig_advcode	orig_origcode
Code Modulation	SNR	-2.8007	21.3578	-2.798	-2.798
	M_dis_∞	16.326	0.945	16.382	16.386
MP3 Compression 64Kb/s	SNR	-2.799	18.582	-2.796	-2.796
	M_dis_∞	16.306	1.489	16.361	16.371
MP3 Compression 96Kb/s	SNR	-2.798	20.866	-2.796	-2.796
	M_dis_∞	16.310	1.026	16.365	16.373
MP3 Compression 128Kb/s	SNR	-2.798	21.213	-2.796	-2.795
	M_dis_∞	16.310	0.973	16.365	16.372

4.2 Overall Performance

In this section, we mainly evaluate the proposed defense approach and explore the properties of audio adversarial examples against ASR systems. To be specific, the performance is evaluated by two metrics: (1) the difference between the original audio waveform, the adversarial audio waveform, and the audio waveform after defense. (2) the efficiency of the proposed mechanism. We compare the two proposed approaches, code modulation and MP3 compression, and calculate the average results for all the audio sampls. For MP3 compression, we set three different compression rates, namely $64Kb/s$, $96Kb/s$, and $128Kb/s$.

Our proposed defense mechanism adopts thresholding to detect the adversarial attack. It counts an audio as an adversarial example if the difference between the original recognition result and the result after filtering is above a threshold, and vice versa. Fig.5 shows the performance of the proposed detection mechanism in terms of ROC curve. To maximise the detection efficiency, we predefine a threshold for each defense approach, and calculate the related False Negative Rate (FNR) and False Positive Rate (FPR), the related precision measurements results are shown in Table.1.

To measure the impact of defense approachs on audios, we compare the differences of audios before and after processing, based on Minkowski distance (Chebyshev distance) and SNR. As shown in Table.2, we list four kinds of comparing groups, the adv_advcode stands for the adversarial audio and the adversarial audio after defense as the first normal reference group, the orig_adv is the group of original audio and the adversarial audio, the orig_advcode means the original radio and the adversarial audio after defense, the last group called orig_origcode, is original audio and original audio after defense as the second normal reference group. These two normal reference groups are designed to measure that how the proposed defense strategy influences normal audios, considering that our filter in the mechanism will filter all the input audios, we need to make sure the clean sample won't be influenced by our mechanism.

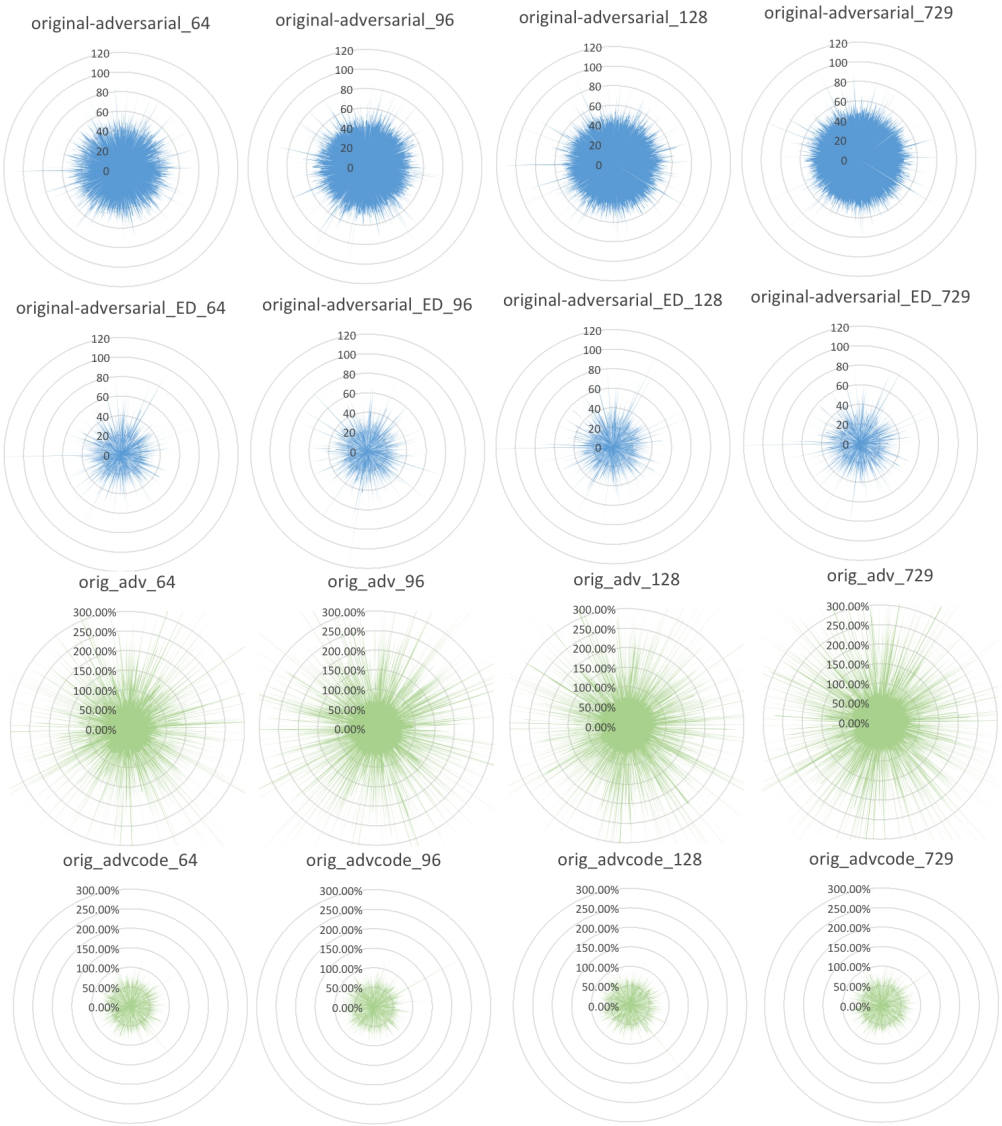

Figure 3: The edit distance of the recognition sentences for attacking and defense. The first three columns are MP3 compression, the number behind (64, 96, and 128) is the compression ratio, and the last column is the code modulation.

The Chebyshev distance shows interesting properties of audio adversarial examples in Table.2. In these four defense approaches, the distances in group orig_adv are the smallest comparing with other groups. Considering about the reason, in the adversarial attacks against ASR systems, the goal of the adversary is to construct the most similar but fooled examples to mislead the system, as defined in Section 2. Therefore, the defense aims at increasing this difference, while being clipped to the normal scope, according the values in the normal reference groups, which can be proved by the distances for orig_advcode group.

As for the SNR evaluation, the group of orig_adv shows the highest values, which means that the noise in the adversarial examples is still obvious; whereas, after being processed by the defense approaches, the differences of SNR values between the original audio and the adversarial audio are much smaller and even closed to the nonmoral reference groups.

A direct way to quantify the defense efficiency is to measure the recognition effect by the protected ASR systems. Thereby, we send the audio adversarial examples and the reconstructed audio examples after defense as the inputs to Deep Speech, and then compare the edit distance of the recognition sentences. Visual distribution of the edit distance between the two groups, original-adversarial group (original audio waveform and adversarial audio waveform) and original-adversarial_ED (original audio waveform and adversarial audio waveform after defense), is illustrated in Fig.3. In this part, two kinds of measurements are deployed, Levenshtein distance and WER edit distance.

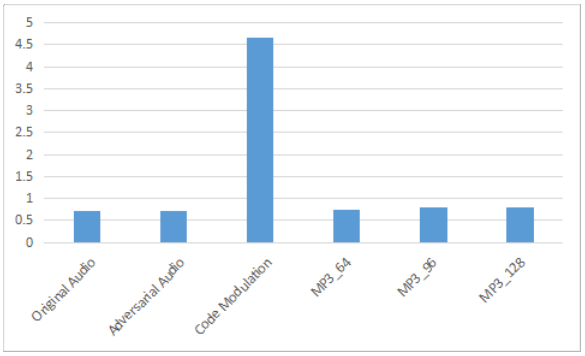

Figure 4: The defense effectiveness evaluated with the recognition time and processing-recognition time (for defense).

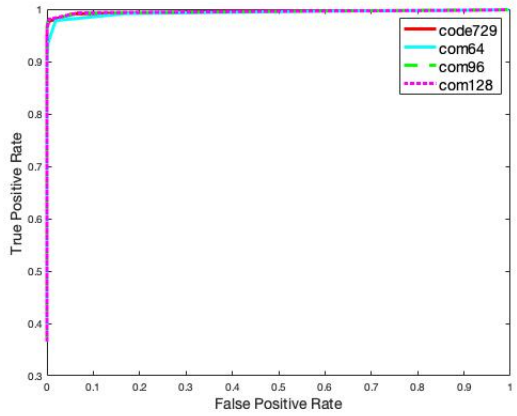

Figure 5: Adversarial attack detection. False Positive Rate is the probability of adversarial audios that are identified as clean audio samples. True Positive Rate measures the percentage of clean audio samples which are correctly identified.

The first line with blue circles shows the difference between the original audio waveform and the adversarial audio, and the second line is the difference between the original audio waveform and the adversarial audio after defense, both measured by the Levenshtein distance. WER edit distance is evaluated in the third and fourth line with the green circles, the former one is the original audio waveform and the adversarial audio, and the latter one is the original audio waveform and the adversarial audio after defense. For both edit distance metrics, the smaller sizes of the circles reflect that our defense approach is much more efficient.

As mentioned before, defense effectiveness is also important when discussing the security problems especially in commercial use. Taking an example of security checking in the airport, the setting of the checkpoint is to increase the security of the airport, but it cannot cause pressure on the normal passage for passengers when guaranteeing the safety of the passenger. Consistent with this situation, we compare the recognition processing time for four types, original audio waveform, adversarial audio waveform and the whole defense and recognition time, the results are shown in Fig.4. We calculate the average time for processing the whole 4070 audio waveforms. The results show that to some extent, the consuming time for defense including the transmitting and recognition just a little longer than recognising the original and adversarial examples, this is reasonable and can be acceptable.

5 RELATED WORK

In the context of two intriguing properties of neural networks, a formal definition of "adversarial examples" was proposed by Szegedy et al. [49]. This was caused by the fact that the input and output mapping of the deep neural work was largely incoherent, the malicious users can make the network incorrectly classify the images by applying some sort of subtle perturbation to maximize the prediction error of the network.

Nonetheless, the work of Goodfellow et al. [20] made a statement that the disturbance of neural networks was caused by their linear characteristics instead of their nonlinearity and overfitting. Furthermore, they proposed a method named fast gradient sign method (FGSM) for finding similar adversarial examples x^* in the L^∞ neighborhood of the original sample x, with the optimization strategy of performing one step gradient update from x in the input space with a volume ϵ. However, this kind of attack was designed to be fast but not to minimize the adversarial perturbations, which shifted all the elements of the input and made easier to be detected by the recently proposed defense approach defense distillation [39], since FGSM may fail to generate successful attacks while other iterative optimization based methods could succeed.

To make the adversarial examples more robust, Carlini and Wagner presented a powerful attack against defensive distillation [9], also known as the *C&W* attack, which generated adversarial examples with smaller perturbations for both targeted and untargeted attacks for three metrics including L^0, L^2 and L^∞, and showed more efficient than other published attacks. Naturally, it became the benchmark and considered as an ideal evaluation method for potential defenses. This is exactly the attack mechanism we defend in this article in the audio field.

Just as mentioned before, generating adversarial examples becomes an critical step to evaluate and improve the robustness of machine learning technologies when dealing with the hard perceptual problems. Over the past two years, many researches have been conducted on the theoretics, application scenarios and defense mechanisms of adversarial examples, an interesting Grundy's game competition situation is presented.

The empirical study of Carlini and Wagner [7] used the L^2 attack algorithm of *C&W* attack to generate targeted adversarial examples and showed that ten recent detection methods could be bypassed by the attackers. Whereas, work in [41] locally corrupted the image by redistributing pixel values via a process termed pixel deflection, which enabled the effective recovery of the true class against a variety of robust attacks. [30] represented a deselection scheme by comparing the action distribution of the current observed frame and the predicted frame from the action-conditioned frame prediction module. Moreover, Shen et al. [44] proposed an effective framework based Generative Adversarial Nets named APE-GAN to defend the

adversarial examples, evaluated on the DenseNet40, InceptionV3 and ResNet50 for MNIST, CIFAR10 and ImageNet.

As the representative work of studying adversarial examples in generative models, Fischer, and Song [27] applied *C&W* attack as the optimization method to generate adversarial examples. They generated adversarial examples for both targeted and untargeted attack against variational autoencoders (VAEs) and VAEs composed with a generative adversarial network (VAE-GANs). In their work, they indirectly manipulated the latent representation and generated a targeted adversarial reconstruction of the input and directly optimized against differences in source and target.

The work in [31] explored the transferability of both targeted and untargeted adversarial examples based on *C&W* attack and other attacks, which showed that the adversarial examples generated to evade some model could mislead other models trained for the same work.

To defend neural network classifers against adversarial examples, Meng and Chen [34] pointed that adding affiliated classifiers can help the protected model classify the normal examples and adversarial examples. In their work, they constructed separate detector networks and a reformer network to learn to differentiate between normal and adversarial examples by approximating the manifold of normal examples. Finally moved adversarial examples towards the manifold of normal examples, which seemed effective for correctly classifying adversarial examples with small perturbation. Lately, Carlini and Wagner [8] stated that the MagNet and the work in [53] were not robust to adversarial examples, and they could construct adversarial examples that defeat these defenses with only a slight increase in distortion based on the *C&W* attack.

Ensemble adversarial training presented in [50] incorporated the disturbed inputs which were transferred from other pre-trained models, exhibit increased robustness to the transferring adversarial examples generated by various single-step and multi-step attacks including *C&W* attack. However, this kind of defense requires a large amount of adversarial examples to model the defense training. Moreover, the adversarial training strategy is specific to certain adversarial examples generating approaches. Considering about the numerous defense approaches, it's normal to wonder whether a strong defense can be created by combining multiple (possibly weak) defenses. However, the work in [24] implied that ensemble of weak defenses is not sufficient to provide strong defense against adversarial examples. As announced in [10, 47], there are still a number of open research challenges in the research of adversarial examples, including the formation mechanism and especially the effective defense.

The aforementioned adversarial examples show us the facts that even the simplest machine learning algorithms, including supervised learning, unsupervised learning and reinforcement learning, can all be attacked and perform unexpected way contrary to the original intention of the designer. Despite the aforementioned generating approaches of adversarial examples in image processing domain. Considering about the huge development and wide applications of audio processing systems, such as the Google Home and Amazon Alexa, recently, Carlini and Wagner [10] constructed the audio adversarial examples on Speech-to-Text systems. They applied white-box iterative optimization-based *C&W* attack to Mozilla's implementation DeepSpeech end-to-end. They turned any audio waveform into any target adversarial transcription with a 100% success rate, which showed a new domain to explore the intriguing properties of neural networks. In this article, we mainly focus on the defense strategies against the attack in [10].

To the best of the authors' knowledge, all the existing defense approaches against adversarial examples mainly focus on the image processing domain, there are few works about defending the audio adversarial examples. Considering about the natural differences between images and audios, especially the statistical redundancy mentioned before, the defense strategies can't be directly applied to deal with the audio adversarial examples. To address this problem, in this article, as the first attempt to mitigate the audio adversarial examples, we propose a new defense strategy for protecting Speech-to-Text systems against adversarial examples.

6 CONCLUSION AND FUTURE WORK

This article explored the properties of adversarial audio examples against Cloud-aided ASR systems, and showed that primitive signal processing transformations may have the potential to defend the adversarial perturbations. Furthermore, we proposed a universal defense against speech recognition systems, including code modulation and audio compression. Code modulation combines the G.729 narrow-band vocoder-based audio data compression algorithm and PCM to transform and convert the audio waveform. Audio compression reduces the residual audio information exceed to human hearing by using MP3 compression. The experimental results showed that our proposed defense approach achieved high performance in terms of the difference between the audio, and the difference between the transformation sequences and the efficiency, based on the metrics of Minkowski distance, SNR, edit distance, and processing time.

The proposed mechanism will inspire more work in defending adversarial attack in audio domain. Furthermore, considering the transferability of adversarial examples in images, the defense strategy should be noticed to deal with the transferability in audio. Additionally, studying the possible behaviors by the attackers and extending the proposed approach to effectively resist the attacks is very crucial, such as gray-box attack and black-box attack depending on the adversary's knowledge about the defense approaches.

REFERENCES

[1] 2012. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Process. Mag.* 29, 6 (2012), 82–97. https://doi.org/10.1109/MSP.2012.2205597

[2] Moustafa Alzantot, Bharathan Balaji, and Mani B. Srivastava. 2018. Did you hear that? Adversarial Examples Against Automatic Speech Recognition. *CoRR* abs/1801.00554 (2018).

[3] Anurag Arnab, Ondrej Miksik, and Philip H. S. Torr. 2017. On the Robustness of Semantic Segmentation Models to Adversarial Attacks. *CoRR* abs/1711.09856 (2017).

[4] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. 2017. Synthesizing Robust Adversarial Examples. *CoRR* abs/1707.07397 (2017).

[5] Vahid Behzadan and Arslan Munir. 2017. Vulnerability of Deep Reinforcement Learning to Policy Induction Attacks. In *Machine Learning and Data Mining in Pattern Recognition - 13th International Conference, MLDM 2017, New York, NY, USA, July 15-20, 2017, Proceedings*. 262–275.

[6] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David A. Wagner, and Wenchao Zhou. 2016. Hidden Voice Commands. In *25th USENIX Security Symposium, USENIX Security 16, Austin, TX, USA, August 10-12, 2016*. 513–530.

[7] Nicholas Carlini and David A. Wagner. 2017. Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods. In *Proceedings of the 10th*

ACM Workshop on Artificial Intelligence and Security, AISec@CCS 2017, Dallas, TX, USA, November 3, 2017. 3–14.

[8] Nicholas Carlini and David A. Wagner. 2017. MagNet and "Efficient Defenses Against Adversarial Attacks" are Not Robust to Adversarial Examples. *CoRR* abs/1711.08478 (2017).

[9] Nicholas Carlini and David A. Wagner. 2017. Towards Evaluating the Robustness of Neural Networks. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*. 39–57.

[10] Nicholas Carlini and David A. Wagner. 2018. Audio Adversarial Examples: Targeted Attacks on Speech-to-Text. In *2018 IEEE Security and Privacy Workshops, SP Workshops 2018, San Francisco, CA, USA, May 24, 2018*. 1–7. https://doi.org/10.1109/SPW.2018.00009

[11] Moustapha Cissé, Yossi Adi, Natalia Neverova, and Joseph Keshet. 2017. Houdini: Fooling Deep Structured Prediction Models. *CoRR* abs/1707.05373 (2017).

[12] Moustapha Cissé, Yossi Adi, Natalia Neverova, and Joseph Keshet. 2017. Houdini: Fooling Deep Structured Visual and Speech Recognition Models with Adversarial Examples. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*. 6980–6990.

[13] Google Cloud. [n.d.]. Google Cloud Speech-to-Text. https://cloud.google.com/speech-to-text/. Accessed March 3, 2019.

[14] Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Li Chen, Michael E. Kounavis, and Duen Horng Chau. 2017. Keeping the Bad Guys Out: Protecting and Vaccinating Deep Learning with JPEG Compression. *CoRR* abs/1705.02900 (2017).

[15] Dialogflow. [n.d.]. Dialogflow. https://dialogflow.com/. Accessed March 3, 2019.

[16] R. Donaldson and D. Chan. 2003. Analysis and Subjective Evaluation of Differential Pulse-Code Modulation Voice Communication Systems. *IEEE Transactions on Communication Technology* 17, 1 (2003), 10–19.

[17] Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M. Roy. 2016. A study of the effect of JPG compression on adversarial images. *CoRR* abs/1608.00853 (2016).

[18] Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. 2017. Robust Physical-World Attacks on Deep Learning Models. *arXiv preprint arXiv:1707.08945* 1 (2017).

[19] Volker Fischer, Mummadi Chaithanya Kumar, Jan Hendrik Metzen, and Thomas Brox. 2017. Adversarial Examples for Semantic Image Segmentation. *CoRR* abs/1703.01101 (2017).

[20] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and Harnessing Adversarial Examples. *CoRR* abs/1412.6572 (2014).

[21] Alex Graves, Santiago Fernández, Faustino J. Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006*. 369–376.

[22] Chuan Guo, Mayank Rana, Moustapha Cissé, and Laurens van der Maaten. 2017. Countering Adversarial Images using Input Transformations. *CoRR* abs/1711.00117 (2017).

[23] Awni Hannun. 2017. Sequence Modeling with CTC. *Distill* 2, 11 (2017), e8.

[24] Warren He, James Wei, Xinyun Chen, Nicholas Carlini, and Dawn Song. 2017. Adversarial Example Defense: Ensembles of Weak Defenses are not Strong. In *11th USENIX Workshop on Offensive Technologies, WOOT 2017, Vancouver, BC, Canada, August 14-15, 2017*.

[25] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. 2017. Query-Efficient Black-box Adversarial Examples. *CoRR* abs/1712.07113 (2017).

[26] Robin Jia and Percy Liang. 2017. Adversarial Examples for Evaluating Reading Comprehension Systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*. 2021–2031.

[27] Jernej Kos, Ian Fischer, and Dawn Song. 2017. Adversarial examples for generative models. *CoRR* abs/1702.06832 (2017).

[28] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. 2016. Adversarial examples in the physical world. *CoRR* abs/1607.02533 (2016).

[29] Zhizhong Li and Derek Hoiem. 2016. Learning Without Forgetting. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*. 614–629.

[30] Yen-Chen Lin, Ming-Yu Liu, Min Sun, and Jia-Bin Huang. 2017. Detecting Adversarial Attacks on Neural Network Policies with Visual Foresight. *CoRR* abs/1710.00814 (2017).

[31] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. 2016. Delving into Transferable Adversarial Examples and Black-box Attacks. *CoRR* abs/1611.02770 (2016).

[32] Shike Mei and Xiaojin Zhu. 2015. The Security of Latent Dirichlet Allocation. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2015, San Diego, California, USA, May 9-12, 2015*.

[33] Shike Mei and Xiaojin Zhu. 2015. Using Machine Teaching to Identify Optimal Training-Set Attacks on Machine Learners. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*. 2871–2877.

[34] Dongyu Meng and Hao Chen. 2017. MagNet: A Two-Pronged Defense against Adversarial Examples. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS 2017, Dallas, TX, USA, October 30 - November 03, 2017*. 135–147.

[35] Tom M. Mitchell, William W. Cohen, Estevam R. Hruschka Jr., Partha Pratim Talukdar, Justin Betteridge, Andrew Carlson, Bhavana Dalvi Mishra, Matthew Gardner, Bryan Kisiel, Jayant Krishnamurthy, Ni Lao, Kathryn Mazaitis, Thahir Mohamed, Ndapandula Nakashole, Emmanouil Antonios Platanios, Alan Ritter, Mehdi Samadi, Burr Settles, Richard C. Wang, Derry Tanti Wijaya, Abhinav Gupta, Xinlei Chen, Abulhair Saparov, Malcolm Greaves, and Joel Welling. 2015. Never-Ending Learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*. 2302–2310.

[36] Mozilla. [n.d.]. Project deepspeech. https://github.com/mozilla/DeepSpeech. 2017.

[37] Anh Mai Nguyen, Jason Yosinski, and Jeff Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. 427–436.

[38] Nicolas Papernot, Patrick D. McDaniel, Ian J. Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. 2016. Practical Black-Box Attacks against Deep Learning Systems using Adversarial Examples. *CoRR* abs/1602.02697 (2016).

[39] Nicolas Papernot, Patrick D. McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. 2016. Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks. In *IEEE Symposium on Security and Privacy, SP 2016, San Jose, CA, USA, May 22-26, 2016*. 582–597.

[40] Sarah Perez. [n.d.]. Microsoft silences its new A.I. bot Tay, after Twitter users teach it racism. https://techcrunch.com/2016/03/24/microsoft-silences-its-new-a-i-bot-tay-after-twitter-users-teach-it-racism/. Accessed Mar 24, 2016.

[41] Aaditya Prakash, Nick Moran, Solomon Garber, Antonella DiLillo, and James A. Storer. 2018. Deflecting Adversarial Attacks with Pixel Deflection. *CoRR* abs/1801.08926 (2018).

[42] ITUT Rec. 1996. G. 729, Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear Prediction (CS-ACELP). *International Telecommunication Union, Geneva* (1996).

[43] Tara N. Sainath, Ron J. Weiss, Kevin W. Wilson, Bo Li, Arun Narayanan, Ehsan Variani, Michiel Bacchiani, Izhak Shafran, Andrew W. Senior, Kean K. Chin, Ananya Misra, and Chanwoo Kim. 2017. Multichannel Signal Processing With Deep Neural Networks for Automatic Speech Recognition. *IEEE/ACM Trans. Audio, Speech & Language Processing* 25, 5 (2017), 965–979. https://doi.org/10.1109/TASLP.2017.2672401

[44] Shiwei Shen, Guoqing Jin, Ke Gao, and Yongdong Zhang. 2017. Ape-gan: Adversarial perturbation elimination with gan. *ICLR Submission, available on OpenReview* (2017).

[45] Daniel L. Silver, Qiang Yang, and Lianghao Li. 2013. Lifelong Machine Learning Systems: Beyond Learning Algorithms. In *Lifelong Machine Learning, Papers from the 2013 AAAI Spring Symposium, Palo Alto, California, USA, March 25-27, 2013*.

[46] SPEECHMATICS. [n.d.]. SPEECHMATICS. https://www.speechmatics.com/. Accessed March 3, 2019.

[47] Ion Stoica, Dawn Song, Raluca Ada Popa, David A. Patterson, Michael W. Mahoney, Randy H. Katz, Anthony D. Joseph, Michael I. Jordan, Joseph M. Hellerstein, Joseph E. Gonzalez, Ken Goldberg, Ali Ghodsi, David Culler, and Pieter Abbeel. 2017. A Berkeley View of Systems Challenges for AI. *CoRR* abs/1712.05855 (2017).

[48] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. 2017. One pixel attack for fooling deep neural networks. *CoRR* abs/1710.08864 (2017).

[49] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *CoRR* abs/1312.6199 (2013).

[50] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Dan Boneh, and Patrick D. McDaniel. 2017. Ensemble Adversarial Training: Attacks and Defenses. *CoRR* abs/1705.07204 (2017).

[51] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. 2014. Deep Domain Confusion: Maximizing for Domain Invariance. *CoRR* abs/1412.3474 (2014).

[52] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan L. Yuille. 2017. Adversarial Examples for Semantic Segmentation and Object Detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. 1378–1387.

[53] Valentina Zantedeschi, Maria-Irina Nicolae, and Ambrish Rawat. 2017. Efficient Defenses Against Adversarial Attacks. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, AISec@CCS 2017, Dallas, TX, USA, November 3, 2017*. 39–49.