Detecting Acoustic BackDoor Transmission of Inaudible Messages Using Deep Learning

Silvija Kokalj-Filipovic
skfilipovic@perspectalabs.com
Perspecta Labs Inc

Morriel Kasher
morrielkasher@gmail.com
Rutgers University

Michael Zhao
azhao2001@gmail.com
Rutgers University

Predrag Spasojevic
spasojev@winlab.rutgers.edu
Rutgers University

ABSTRACT

The novel secret inaudible acoustic communication channel [11], referred to as the BackDoor channel, is a method of embedding inaudible signals in acoustic data that is likely to be processed by a trained deep neural net. In this paper we perform preliminary studies of the detectability of such a communication channel by deep learning algorithms that are trained on the original acoustic data used for such a secret exploit. The BackDoor channel embeds inaudible messages by modulating them with a sinewave of $40kHz$ and transmitting using ultrasonic speakers. The received composite signal is used to generate the Backdoor dataset for evaluation of our neural net. The audible samples are played back and recorded as a baseline dataset for training. The Backdoor dataset is used to evaluate the impact that the BackDoor channel has on the classification of the acoustic data, and we show that the accuracy of the classifier is degraded. The degradation depends on the type of deep classifier and it appears to impact less the classifiers that are trained using autoencoders. We also propose statistics that can be used to detect the out-of-distribution samples created as a result of the BackDoor channel, such as the log likelihood of the variational autoencoder used to pre-train the classifier or the empirical entropy of the classifier's output layer. The preliminary results presented in this paper indicate that the use of deep learning classifiers as detectors of the BackDoor secret channel merits further research.

CCS CONCEPTS

• **Security and privacy → Intrusion detection systems**; • **Hardware → Sound-based input / output**; **Digital signal processing**; • **Computing methodologies → Machine learning**; **Neural networks**; Learning latent representations;

KEYWORDS

ultrasonic acoustics, neural networks, deep learning, inaudible voice commands, BackDoor channel, ultrasound injection

WiseML '20, July 13, 2020, Linz (Virtual Event), Austria

ACM ISBN 978-1-4503-8007-2/20/07…$15.00
https://doi.org/10.1145/3395352.3402629

ACM Reference Format:
Silvija Kokalj-Filipovic, Morriel Kasher, Michael Zhao, and Predrag Spasojevic. 2020. Detecting Acoustic BackDoor Transmission of Inaudible Messages Using Deep Learning. In *2nd ACM Workshop on Wireless Security and Machine Learning (WiseML '20), July 13, 2020, Linz (Virtual Event), Austria.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3395352.3402629

1 INTRODUCTION AND PROBLEM STATEMENT

The BackDoor ultrasonic exploit by Roy et al. [11] allows for playing of inaudible signals at frequencies above the human hearing threshold that can be recorded by an unmodified receiver microphone as if they were played audibly. The unrelated BackDoor attack on Deep Learning (DL) systems by Chen et al. [2] is a form of poisoning attack on deep learning systems. In contrast, the BackDoor ultrasonic exploit [11] forms a secret secondary channel for the transmission of acoustic data that may not be related to DL. This approach requires only a simple ultrasonic transmitter to send the secret audio intended to be decoded as voice that is legitimately aired. Hence, its natural application is machine-to-machine acoustic communications, typically processed by the DL based automatic speech recognition (ASR) methods. Our aim is to ascertain whether the backdoor ultrasonic communication channel is easily detectable with existing deep learning algorithms, such as those used by ASR.

The applications of deep learning in the audio domain are copious and very diverse with, frequently, very complex models (e.g., for speech-to-text translation). In order to avoid complex and overly large models we opted to demonstrate our BackDoor detection approach using a simple application of keyword spotting that aims to recognize command words. We use our own neural networks for classification although other neural networks may achieve better classification results. The main reason is familiarity: we successfully tested these for robustness against adversarial attacks on the same speech commands used in the datasets here [7]. Upon analyzing spectral components of the Backdoor recordings, which when played back could not be distinguished from the regularly delivered audio, we hypothesized that the effect of the BackDoor voice delivery on a DL classifier may be similar to the effect of adversarial attacks in that minimal perturbations of the original datapoints can cause misclassification [4]. In fact, Backdoor signals processed by the DL classifier can be treated as adversarial attacks in the physical world [10, 15].

In the Speech-to-Text domain the seminal research by N. Carlini and D. Wagner has shown that the adversarial example does not persist Over The Air (OTA) [1]. This work opened a question whether

there exist generic, transferable (robust) ways to produce audio that sounds like noise to humans, sounds like a valid command to an ASR system and works against both speech and speaker recognition systems. It was later shown that if the adversarial speech is embedded into music, then the song as a carrier of the adversarial example can deliver the attack OTA, as illustrated in the Commander Song approach [16].

In the seminal paper [8], the authors define adversarial examples for voice commands as consisting of a recording that seems to be innocuous to a human observer (such as a song or speech), but contains voice commands recognized by a machine learning algorithm. This oddly sounds like the setup that we are presenting here, which can be misleading since the BackDoor channel has not been designed to be recognized by the machine learning algorithm in the manner of a targeted attack. Rather, the design goal was auditory imperceptibility. Also, the synthesis of the Backdoor samples is hardware specific. On the other hand, as the exploited hardware is pervasive, if such an unintended attack is possible, it certainly deserves attention from both practical and theoretical standpoints. The machines that synthesize and receive speech commands (such as robotic communication) are especially vulnerable, as the underlying technology is always machine learning and as it is easier to modify synthesized speech than the human voice commands. We here perform initial investigation on the ability of machine learning algorithms to distinguish voice commands from those delivered via an embedded inaudible acoustic communication channel.

Problem statement through a usage scenario: We imagine a scenario in which a receiving device receives ordinary voice commands from a transmitting device at short-to-medium range through the standard acoustic channel played over the air. An attacker modifies the transmitting device such that it is equipped with ultrasonic speakers that inaudibly modify the original audio by multiplexing false commands. Human auditory system can only detect the original audio.

As previously mentioned, it is quite possible that such an exploit could use machine-to-machine (M2M) acoustic communication to piggyback the ultrasonic message. M2M (robotic) acoustic communication is typically processed by the DL based ASR. For example, the attacker may want to send a 'Stop' command through the secret inaudible ultrasonic transmission and cause the robotic receiver to shut-down. Another target of the Backdoor delivery can be a speaker recognition system (perhaps distinguishing between male and female speakers), where the inaudible channel can impersonate speakers that are not present affecting the decision making based on the classifier. The machines that receive the signal and process it automatically are not usually equipped with the signal processing software that can detect usage of the BackDoor channel. If a detector existed that could distinguish the BackDoor data sample even when the classifier does not, the attacker can also compromise the machine by injecting the detector code and have it filter out the sequence of inaudible messages thus creating a real secret communication channel. We seek to investigate if the receiver machines equipped with a DL neural net used for legitimate ASR can detect that the received signal is compromised by BackDoor. We hope to then use it to identify if the voice signals from the transmitter are just the original audible channel or the BackDoor channel transmissions.

2 BACKDOOR SYSTEM

2.1 Acoustic Systems Background

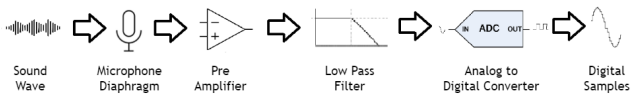

Figure 1: Standard Acoustic Receiver Design

Acoustic receiver design for consumer microphones is largely standardized, maintaining similar construction among a variety of devices as illustrated in Fig. 1. First, the air pressure waves that compose sound reach the microphone, whereupon they actuate its diaphragm. The diaphragm is a thin membrane that vibrates in response to this pressure variation, and in doing so creates an electrical signal on the order of tens of milliVolts. This signal requires amplification for further processing and sampling, and thus goes through a pre-amplifier (or pre-amp) with a gain of around ten. Since most analog-to-digital converters (ADC) in microphones operate at $\approx 48kHz$, the resultant amplified signal is first filtered through a low-pass filter (LPF) with a cutoff frequency around $24kHz$ (to keep it below the Nyquist frequency) before finally being converted to digital samples by the ADC. This sampling rate is appropriate because the human audible hearing range is restricted to between $20Hz$ and $20kHz$, with some infants and animals being capable of hearing up to $24kHz$.

An important characteristic of microphones is the operation of their pre-amp stage. It is designed such that the signal from the diaphragm should be linearly amplified by some gain factor $S_{out} = A_1 S_{in}$, where A_1 is typically ≈ 10.

In practice, however, imperfections in the design and construction of the pre-amp (often due to cost-saving measures) result in the signal being amplified by a non-linear series of the form

$$S_{out} = \sum_{k=1}^{\infty} A_k (S_{in})^k = A_1 S_{in} + A_2 (S_{in})^2 + A_3 (S_{in})^3 + \dots \quad (1)$$

Each A_k coefficient is weaker than the previous, such that cubic terms (A_3) and above are negligible and largely indistinguishable above the noise floor. For frequencies in the typical range (below $24kHz$), the quadratic term (A_2) is also generally weak enough to be unnoticeable. However, this quadratic term becomes non-negligible for frequencies beyond the audible range (above $24kHz$). This is the source of the BackDoor exploit developed by Roy et al. [11] for the inaudible transmission and recording of sound.

2.2 BackDoor Exploit Description

In order to leverage the non-linearity in the pre-amp of microphones, we begin by creating a signal of the form $\sin(f_c 2\pi t)$, in which a carrier frequency (denoted f_c) is greater than $24kHz$. Transmitting this will produce an inaudible signal that will be filtered out by the built-in LPF of the receiver microphone. However, by simultaneously transmitting another signal of the form $\sin(f_s 2\pi t)$ in which a secondary frequency (denoted f_s) is also greater than $24kHz$, the resultant signal reaching the receiver microphone can be modeled as the sum of both signals such that

$$S_{in}(t) = \sin(f_c 2\pi t) + \sin(f_s 2\pi t) \quad (2)$$

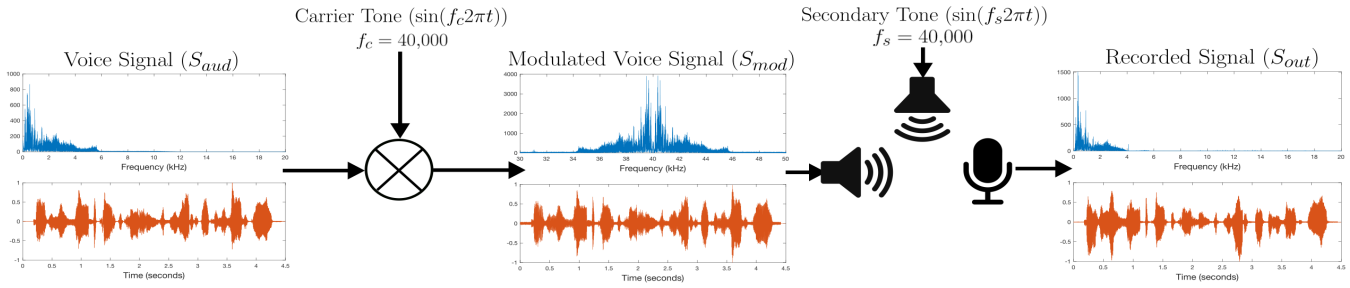

Figure 2: BackDoor Example for Voice Sample

Upon reaching the pre-amp stage of the microphone the signal will be non-linearly amplified such that the resultant signal will be of the form

$$S_{out}(t) \approx A_1(S_{in}) + A_2(S_{in})^2 = A_1\left[\sin(f_c 2\pi t) + \sin(f_s 2\pi t)\right] + A_2\Big[1 + \cos((f_c - f_s)2\pi t) + \cos((f_c + f_s)2\pi t) + \frac{1}{2}\left(\cos(2f_c 2\pi t) + \cos(2f_s 2\pi t)\right)\Big] \quad (3)$$

After passing through the LPF, all frequencies above $24kHz$ are filtered out, and the only component of $S_{out}(t)$ that remains is the $A_2\cos((f_c - f_s)2\pi t)$ term. Thus, the ADC will produce digital recorded samples of a tone with frequency equal to the difference between the frequencies of the carrier and secondary tones ($f_c - f_s$). By selecting f_c and f_s appropriately, both can be inaudible while played through the air but their difference can map to an audible frequency ($f_c, f_s > 24kHz; 20Hz < f_c - f_s < 20kHz$) and thus be recorded as if it was played audibly. While this method is effective in inaudibly transmitting and recording tones of a single frequency, the goal of such inaudible communication is to record entire samples of audio, particularly that of human speech. To do this, further processing is needed prior to transmission. Song et al. [13] developed a technique using amplitude modulation to accomplish this. Given some original audio pattern $S_{aud}(t)$, we create an ultrasonic signal by modulating the amplitude of our carrier tone such that the new signal is of the form $S_{mod}(t) = S_{aud}(t)\sin(f_c 2\pi t)$. A secondary sinewave with $f_s = f_c$ is transmitted simultaneously. The signal received at microphone is

$$S_{in}(t) = S_{aud}(t)\sin(f_c 2\pi t) + \sin(f_c 2\pi t) \quad (4)$$

where it is linearly and quadratically amplified by the non-linearity of the microphone's pre-amp. The result after low-pass filtering will then be of the form

$$S_{out}(t) = \frac{A_2}{2}\left(1 + 2S_{aud}(t) + S^2_{aud}(t)\right) \quad (5)$$

$S^2_{aud}(t)$ is significantly lower than $S_{aud}(t)$ above DC frequencies, and hence, since most human speech patterns contain only frequencies above $50Hz$, its effect can be neglected.

Fig. 2 illustrates the entire BackDoor process for a sample voice signal. The signal is modulated with a $40kHz$ carrier tone, creating a signal with sidebands around a $40kHz$ center frequency (where the lower sideband is still above the audible threshold for humans). When this is played through an ultrasonic speaker at the same time as a $40kHz$ secondary tone, the microphone records a signal $S_{out}(t)$ that is almost identical to the original signal $S_{aud}(t)$. The effective result is that $S_{aud}(t)$ is recorded by the receiver microphone while remaining inaudible throughout its transmission. This can be applied for the manipulation of Voice-Enabled Devices (VED) such as speech-activated personal assistants, as demonstrated by Zhang et al. [17]. Such an exploit permits an attacker to prompt the VED to discreetly perform privileged actions or to divulge personal information. This presents a need for the detection of such BackDoor-transmitted inaudible voice commands to mitigate this security risk.

Previous research by Roy et al. [12] has primarily focused on detecting BackDoor using signal processing based on its nuanced noise characteristics. These include observing the prevalence of tones below $50Hz$ (owing to the existence of the above $S^2_{aud}(t)$ term) and applying angle-of-arrival-based techniques to estimate the phase-based separation of the two ultrasonic speakers. However, these are susceptible to interference from slight environmental changes and can be unreliable for robust attacks or attacks of varying power.

2.3 BackDoor Implementation

Our implementation of the BackDoor system begins with the input of an original audio recording as a wave file. This file is then upsampled to $250kHz$ so that it can be modulated with the $40kHz$ carrier tone in software. The result is then amplified by a factor N_1 before being converted to an analog signal by a National Instruments myDAQ Data Acquisition Device [6]. The output from the myDAQ is insufficient to drive the ultrasonic speakers used, so a driver circuit was implemented for each of our two ultrasonic speakers using NE5534AP-based non-inverting operational amplifiers powered by isolated power supplies. The speakers themselves were Prowave 400ST160 ultrasonic transducers [3]. The frequency response of these speakers indicates that their ideal resonant frequency is $40kHz$, which is why this was chosen as the carrier tone f_c. A $40kHz$ secondary tone (f_s) with amplitude 1 V_{pp} was produced by a Tektronix AFG3021 Function Generator [9] and amplified by a gain factor $A_V = 2$ in our driver circuit before being played through the secondary speaker. Experimentation indicated that ideal results were produced with S_{mod} that has twice the amplitude of the secondary tone, so N_1 was chosen to be 2 and the signal produced from the myDAQ was again amplified by our op-amp driver by a gain factor of $A_V = 2$.

Data Collection Procedure: Each sample was played audibly through the air using a standard 4-ohm speaker that was driven using the same driver circuitry as the ultrasonic speakers with a 0.3 V_{pp} input from the myDAQ and a gain factor of $A_V = 2$. All

samples that were played audibly were recorded inside a box with dimensions 0.4m x 0.3m x 0.1m at a distance 0.2m away from a receiver microphone. Next, all samples were played through the BackDoor system described previously at the same distance with the two ultrasonic speakers side-by-side. By comparing the audible re-recorded samples to the BackDoor samples, the confounding variable of background noise was controlled for and thus the algorithm could be appropriately trained to detect BackDoor signals from audible signals of the same type. The recordings were made and processed in $16kHz$ 32-bit audio.

3 DEEP LEARNING

In this subsection we describe the acoustic datasets and the deep learning architectures used to asses the effect of the BackDoor channel on the dataset. We will use standard datasets, which are collections of wave files, prepared for the training and testing of diverse ASR algorithms. We then describe how features are prepared for specific deep learning algorithms starting from those file collections.

3.1 Dataset Sources and Feature Preprocessing

We used two audio datasets to embed BackDoor channels:

- *Speech Command (SC)* dataset [14], which consists of spoken words designed to help train and evaluate keyword spotting systems. Speech Commands data set was collected by Google and released under a CC BY 4.0 license. The dataset is labeled by *Speech words*.
- *Audio Set* dataset [5], which contains samples from human speech uploaded to YouTube voluntarily by a variety of speakers. Audio Set data set was collected by Google and released under a CC BY-SA 4.0 license. The dataset is labeled by the *Speaker gender*.

The SC dataset is an attempt to build a standard training and evaluation dataset for a class of simple speech recognition tasks. Keyword spotting aims to build simple and effective models for identification of keywords in utterances, particularly for the purpose of deploying such algorithms on mobile platforms: a familiar example would be iPhone and Alexa devices recognizing the 'wake-up' commands 'Siri' and 'Alexa', respectively. We use this dataset to demonstrate the concept of using the proposed deep learning metrics to detect the inaudible BackDoor channel embedded in the legitimate utterances that are effectively classified by the associate deep learning algorithm. We present the keyword spotting classification on just two keywords from the dataset (Wow and Stop) in order to compare the results with another binary dataset containing two classes of speech — by male and female speakers. The other dataset, Audio Set, contains 30 samples (7-12 seconds each), of which 15 are classified as male and 15 are classified as female. We created two datasets from each of the above datasets: the audible version of each sample (played through a speaker and re-recorded), and the BackDoor version of each sample (played ultrasonically and recorded through BackDoor).

To explain how the dataset is being preprocessed and fed into the deep neural net, we will use the SC dataset, which contains one wave file per utterance and its associated label. The duration of each utterance is one second. For each wave file we perform the standard method of preprocessing audio recordings for speech recognition, except for the last step. Note that wave files are digital formats of audio recordings, obtained by sampling the analog audio signal. For the purposes of evaluating the BackDoor detection, we play back the original SC wave files and record them, and we play them back again while embedding them in the ultrasonic carrier to render them inaudible, and record the received signal into wave files that will be used to build datapoints of the BackDoor dataset. The sampling rate (the number of samples per second) used for both datasets was $16kHz$ (a little stronger than Nyquist criterion, hence without distortions and with full information content about the signals). Therefore, the length of the array of numbers that we get as a result of reading each wave file is 16000. Now, such an array is sparse, and therefore typically the approach is to create a spectrogram from those samples, which performs dimensionality reduction and a convenient visualization of where the voice energy is in the frequency and the time.

The following explains how this is done. We will use 40 features (frequency bands) for each temporal window of 20 milliseconds that slides over the 1s long recording of the keyword in the wave file. 40 features yield decent quality of reproduction. The shift between windows is another parameter, in this case set up so that there are 101 windows covering the utterance of the word. Each window is where a Short Time Fast Fourier Transform (STFFT) is performed, providing the frequency signature. As each window produces the signature with 40 features related to how the energy of the voice is distributed along the frequencies, the obtained spectrogram (visualized utterance) is a 2-D matrix 40 by 101, whose elements are color coded "'FFT coefficients"'. For voice processing there is an additional filter applied to make the spectrograms: the 40 features (frequency coefficients per time instant) are obtained by non-uniformly sampling the audio frequencies according to the human audio perception. Those frequencies are known as Mel frequencies, and the entire transform is known as Mel-Frequency Cepstrum (MFC) transform. In addition, a log operation is performed over the elements of the spectrogram to perform the processing in the dB domain (this is a frequent measure of the relative intensity of audio signals).

Bottom line is that the set of preprocessed features now has 4040 (40 by 101) dimensions per data point instead of 16000. The last step in our preprocessing of Speech commands is different from the typical approach. We flatten the spectrogram by unfolding it, and hence, our input data point is a vector of 4040 elements. This is done for simplicity. Consequently, the convolutional autoencoder and the matching classifier described in the next subsection use 1-D convolutional layers instead of 2-D convolutions.

3.2 Deep Neural Networks and Metrics

We here propose two types of deep learning classifiers for speech signals:

Pretrained by Autoencoder (AE): Inspired by our work on mitigating adversarial examples for both acoustic and radio frequency waveforms, we hypothesized that the AE-based training increases the classifier's robustness to the BackDoor channel in the same way as it protects against adversarial examples, by projecting datapoints to a lower-dimensional manifold (the bottleneck layer of

the AE). This projection is achieved by training the AE to minimize the distance between the original datapoints at the input and their estimates at the output of this particular DL architecture consisting of an encoder and a decoder linked by the bottleneck layer. The upper part of the AE, the encoder, is then retrained in a supervised manner to classify datapoints according to the classes in the labeled training dataset. If the hypothesis is confirmed, such robustness to the BackDoor channel would here be an impediment to its detection, as the main goal is to catch the secret communication and not to make it imperceptible to the classifier. However, our results (Fig. 3 and 4) show that even if the accuracy is well preserved, the DL statistics still provide an indication that the sample is Backdoor delivered.

Classically trained Convolutional Neural Net (CNN): We also use a classically trained CNN that has the same architecture as the encoder of the AE (modulo the last softmax layer used for the classification). We hypothesize that this CNN will experience performance degradation in terms of accuracy when presented with the datapoints that embed the BackDoor channel. If the hypothesis is proved, such degradation due to the BackDoor channel would indicate that we can derive good detection statistics based on the CNN parameters.

Note that we cannot use the accuracy as a detection statistic, as we do not have the ground truth when performing the hypothesis testing on a data sample. As we here know the ground truth for our datasets, both original and those with the embedded BackDoor channel, we can use the accuracy as an indication of the effect that the BackDoor channel has on the classifier. However, for detection we need a different metric. We propose to use the likelihood loss calculated for a variational encoder with the same architecture as the AE, but which treats the bottleneck layer as multivariate Gaussian parameters of the latent distribution to be trained, and performs sampling from that latent distribution every time it pushes the data from encoder to decoder. Similarly the output of the decoder is treated as multivariate Gaussian parameters of the data likelihood, from which we can calculate log likelihood for each particular datapoint. It is expected that out-of-distribution samples, such as the audio sample with the embedded BackDoor, will have a measurably different log likelihood (see Fig. 4). For the classically trained classifier we use the empirical entropy of the softmax layer pseudo-probabilities (see Fig. 5):

$$P(y=c|x) = \frac{\exp(w_c^T f_s(x) + b_c)}{\sum_{i \in C} \exp(w_i f_s(x) + b_i)}, \tag{6}$$

where the class label $c \in C$ is here binary, and f_s represents the layers above the final layer. The architecture of the encoder can be described as repetitive layering of the following substructure of layers that we refer to as *E cell*: $E_C(k_s) = 1DC(k_s) \longrightarrow MaxPooling(2)$, where in each 1-dimensional CNN layer $1DC$ we used 8 channels, and k_s denotes the 1-dimensional kernel size. This created the following encoder design:

$$\begin{aligned} E = E_C(10) &\longrightarrow E_C(5) \longrightarrow E_C(5) \longrightarrow E_C(3) \\ &\longrightarrow 1DC(3) \longrightarrow FC(505) \longrightarrow FC(105), \end{aligned} \tag{7}$$

where $FC(n)$ denotes a fully connected layer with n outputs. All E layers have the same ReLU nonlinearity. For the classification we add to the pretrained encoder another fully connected layer with softmax nonlinearity and 2 outputs, hence creating the following architecture of the AE-pretrained classifier: $C_{AE} = E \longrightarrow FC(2)$. For the pre-training of the encoder we embed it in an autoencoder (AE) architecture together with the decoder D, $AE = E \longrightarrow D$, where D can be described as repetitive layering of a substructure of layers that we refer to as *D cell*: $D_C(k_s) = 1DC(k_s) \longrightarrow UpSampling(2)$, where the upsampling layer performs the operation opposite of max pooling. This created the following decoder design

$$\begin{aligned} D = D_C(3) &\longrightarrow D_C(3) \longrightarrow D_C(5) \\ &\longrightarrow 1DC(5) \longrightarrow 1DC(10), \end{aligned} \tag{8}$$

where all layers have the same ReLU nonlinearity. The classically trained classifier shares the architecture of C_{AE}, only without AE-pretraining the E-layers.

4 EVALUATION RESULTS

We now illustrate the impact that the BackDoor channel has on two types of classifiers, the AE- and the classically trained. We present this impact visually through graphs as they illustrate the trends well, and also because the rigorous presentation of the test metrics requires much more space. The illustrations combine the effects on the accuracy and the effects on the loss functions that indicate out-of-distribution samples. Accuracy results for the two applications (speaker, and speech command recognition) are presented by Figures 3 and 6, respectively. Fig. 3 represents typical (average) results for the speaker gender recognition, which aligns with our expectations. The performance of both classifiers is degraded when the samples are recorded through a BackDoor channel. As expected, the AE-trained classifier is more robust (dashed-line plots). However, with the SC dataset we see the AE-trained classifier behaving both ways: Fig. 6 shows the case when the BackDoor testing subsets performed better on the classically trained classifier. This requires more testing.

Fig. 4 illustrates that the log-likelihood based loss can be used as a statistic for detecting the BackDoor generated outliers. In Fig. 4 the log likelihood during AE training (blue and orange), and during the inference (green and red), are shown for the legitimate and BackDoor-corrupted data, respectively. Equivalently, the pseudo-probabilities at the outputs of the two classifiers shown in Fig. 5 can be used to calculate pseudo-entropy statistics, $H(x) = \sum_{i=1}^{C} p_i(x) \log p_i(x)$, where C is the number of classes, and $[p_1(x) \cdots, p_C(x)]$ is the vector of the class probabilities at the output layer, evaluated at input x. Class probabilities in Fig. 5 are color-coded: yellow for male (M), and purple for female (F). It is obvious that the AE keeps the pseudo-probabilities closer to the ideal points, like on the right plots, i.e., $p(M) = (0, 1)$ and $p(F) = (1, 0)$, making the classes linearly separable, while the classical net moves the pseudo-probabilities closer to the uncertainty area around $(0.5, 0.5)$.

Finally, in Fig. 7 we show the effect of the BackDoor channel on the confusion matrices of both binary classifiers (here for the speech command recognition).

5 CONCLUSION

In this paper we study the effect that a secret inaudible acoustic communication channel, referred to as the BackDoor channel, has on the classification of its embedded acoustic data, and show that

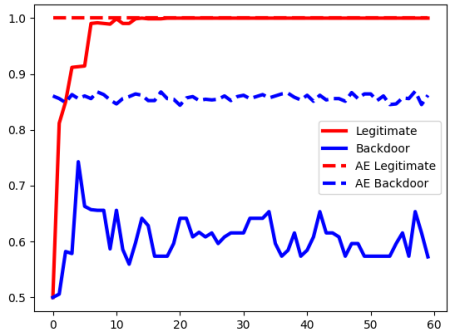

Figure 3: Accuracy of Deep Learning classifiers trained on Male/Female Speaker recognition (over training epochs)

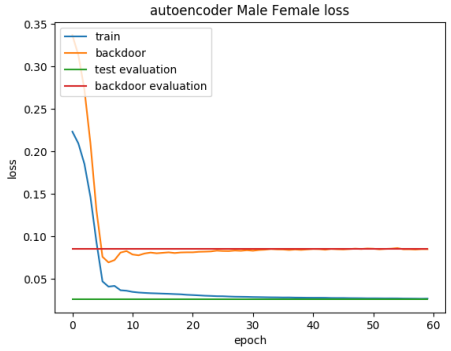

Figure 4: Normalized Log likelihood Loss of Deep Learning classifiers trained on Male/Female Speaker recognition

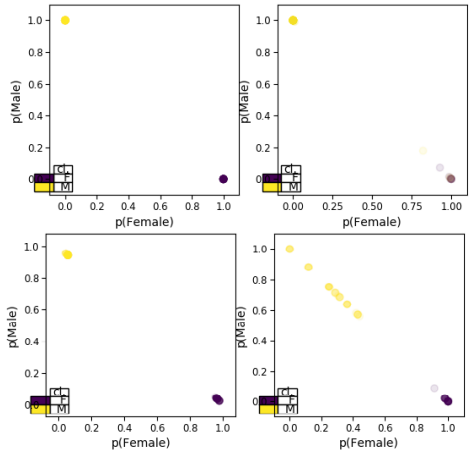

Figure 5: The probabilities of classes derived from the output layer of the Male/Female Speaker recognition (AE top, classical bottom, normal right, BackDoor left)

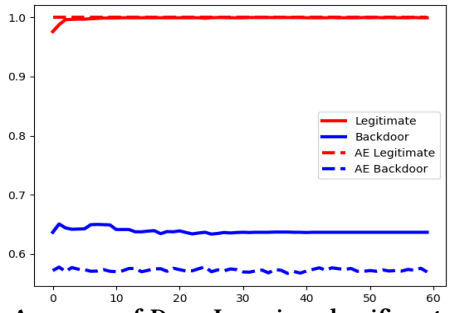

Figure 6: Accuracy of Deep Learning classifiers trained on Speech Commands (over training epochs)

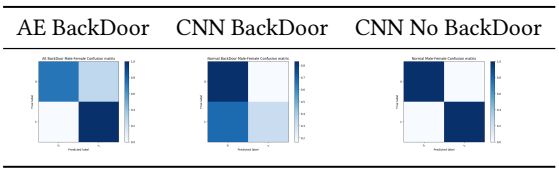

Figure 7: Confusion matrices of Deep Learning classifiers trained on Speech Commands with and without BackDoor

the accuracy of a deep learning (DL) classifier is degraded due to the secret inaudible delivery of the signal. The degradation depends on the type of the DL classifier and it appears to impact less the classifiers that are trained using autoencoders. We perform preliminary studies on the detectability of the BackDoor channel by DL algorithms that are trained on the original data. We propose test statistics that can be used to detect BackDoor even when the accuracy of classification does not indicate it: the log likelihood of the variational autoencoder used to pretrain the classifier, and the empirical entropy of the classifier's output layer. The results indicate that there is a strong case for using DL neural nets to detect inaudible BackDoor communications, but the exact nature of this relationship necessitates further research.

REFERENCES

[1] Nicholas Carlini and David Wagner. 2018. Audio adversarial examples: Targeted attacks on speech-to-text. In *Deep Learning and Security Workshop*.

[2] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Xiaodong Song. 2017. Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning. *ArXiv* abs/1712.05526 (2017).

[3] S. Square Enterprise Company Limited Pro-Wave Electronics Corporation. 2019. *Air Ultrasonic Ceramic Transducers 400ST/R160.* Retrieved May 8, 2020 from http://www.farnell.com/datasheets/1686089.pdf?_ga=2.256607115.1881374495.1588917674-2094016181.1588917674

[4] I. J Goodfellow, J. Shlens, and C. Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint* (2014).

[5] Google Inc. 2017. *Audioset Ontology Human Speech.* Retrieved May 9, 2020 from https://research.google.com/audioset///ontology/speech.html

[6] National Instruments. 2019. *NI myDAQ Device Specifications.* Retrieved May 8, 2020 from https://www.ni.com/pdf/manuals/373061g.pdf

[7] S. Kokalj-Filipovic, R. Miller, and G. Vanhoy. 2019. Adversarial Examples in RF Deep Learning: Detection and Physical Robustness. In *IEEE Global Conf. on Signal and Inform. Processing (GlobalSIP).*

[8] A. Kurakin, I. Goodfellow, and S. Bengio. 2016. Adversarial examples in the physical world. *arXiv preprint* (2016).

[9] Test Equipment Solutions Ltd. 2019. *Arbitrary/Function Generators.* Retrieved May 8, 2020 from http://www.testequipmenthq.com/datasheets/TEKTRONIX-AFG3021-Datasheet.pdf

[10] Yao Qin, Nicholas Carlini, Ian J. Goodfellow, Garrison W. Cottrell, and Colin Raffel. 2019. Imperceptible, Robust, and Targeted Adversarial Examples for Automatic Speech Recognition. In *ICML.*

[11] Nirupam Roy, Haitham Hassanieh, and Roy Romit Choudhury. 2017. BackDoor: Making Microphones Hear Inaudible Sounds. *MobiSys*, Article 5 (June 2017). https://doi.org/10.1145/3081333.3081366

[12] Nirupam Roy, Sheng Shen, Haitham Hassanieh, and Romit Roy Choudhury. 2018. Inaudible Voice Commands: The Long-Range Attack and Defense. In *15th USENIX Symposium on Networked Systems Design and Implementation (NSDI 18).* USENIX Association, Renton, WA, 547–560. https://www.usenix.org/conference/nsdi18/presentation/roy

[13] Liwei Song and Prateek Mittal. 2017. POSTER: Inaudible Voice Commands. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (Dallas, Texas, USA) *(CCS '17).* Association for Computing Machinery, New York, NY, USA, 2583–2585. https://doi.org/10.1145/3133956.3138836

[14] Pete Warden. 2018. Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition. *ArXiv* abs/1804.03209 (2018).

[15] Hiromu Yakura and Jun Sakuma. 2019. Robust Audio Adversarial Example for a Physical Attack. In *IJCAI.*

[16] Yuan, Xuejing et al. 2018. Commandersong: A Systematic Approach for Practical Adversarial Voice Recognition. In *the 27th USENIX Conf. on Security.*

[17] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. 2017. DolphinAttack: Inaudible Voice Commands. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (Dallas, Texas, USA) *(CCS '17).* Association for Computing Machinery, New York, NY, USA, 103–117. https://doi.org/10.1145/3133956.3134052