Investigating a Spectral Deception Loss Metric for Training Machine Learning-based Evasion Attacks

Matthew DelVecchio, Vanessa Arndorfer, William C. Headley
Hume Center for National Security and Technology, Virginia Tech
[matdd96,arndorvf,cheadley]@vt.edu

ABSTRACT

Adversarial evasion attacks have been very successful in causing poor performance in a wide variety of machine learning applications. One such application is radio frequency spectrum sensing. While evasion attacks have proven particularly successful in this area, they have done so at the detriment of the signal's intended purpose. More specifically, for real-world applications of interest, the resulting perturbed signal that is transmitted to evade an eavesdropper must not deviate far from the original signal, less the intended information is destroyed. Recent work by the authors and others has demonstrated an attack framework that allows for intelligent balancing between these conflicting goals of evasion and communication. However, while these methodologies consider creating adversarial signals that minimize communications degradation, they have been shown to do so at the expense of the spectral shape of the signal. This opens the adversarial signal up to defenses at the eavesdropper such as filtering, which could render the attack ineffective. To remedy this, this work introduces a new spectral deception loss metric that can be implemented during the training process to force the spectral shape to be more in-line with the original signal. As an initial proof of concept, a variety of methods are presented that provide a starting point for this proposed loss. Through performance analysis, it is shown that these techniques are effective in controlling the shape of the adversarial signal.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; *Supervised learning by classification.*

KEYWORDS

adversarial machine learning, cognitive radio security, radio frequency machine learning

ACM Reference Format:
Matthew DelVecchio, Vanessa Arndorfer, William C. Headley. 2020. Investigating a Spectral Deception Loss Metric for Training Machine Learning-based Evasion Attacks. 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/10.1145/3395352.3402624

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

ACM ISBN 978-1-4503-8007-2/20/07…$15.00
https://doi.org/10.1145/3395352.3402624

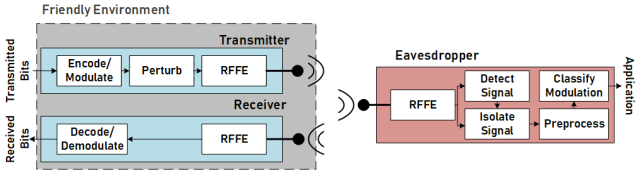

Figure 1: A wireless communications scenario in which an intended communications link is being eavesdropped by a machine learning based spectrum sensor. The transmitter utilizes an adversarial evasion attack to intelligently "perturb" its signal to evade the eavesdropper [4].

1 INTRODUCTION

Boosted by continued improvements in areas such as processing power, storage capacity, and architectural improvements, machine learning algorithms have seen increased usage in recent years and have shown great potential benefit in a wide variety of research fields. One such emerging research field is signal processing, where research has focused on utilizing recent advancements in machine learning to improve on traditional digital signal processing techniques through increased performance and/or a reduced need for *a priori* knowledge. Example signal processing applications showing promise in their utilization of machine learning include spectrum signal detection, synthetic modulation schemes, direction of arrival calculation, jamming detection [1, 7], and automatic modulation classifiers (AMC) [17, 18, 21, 22], among many others. AMC research in particular has shown significant promise in utilizing machine learning to reduce requirements on pre-defined expert features by utilizing state-of-the-art convolutional neural networks for performing both feature extraction and classification tasks [11, 12].

Given the improvements that machine learning offers, and thus the adoption of such methods in real world applications, the security of these networks must be further considered. Recent research has shown that adversarial attacks can harm the performance of machine learning networks by forcing misclassifications or otherwise causing the network to operate in ways orthogonal to its intended use or desired application [8, 13, 19]. Various adversarial techniques can be used to attack a machine learning network, such as poisoning [20], Trojan [3], and evasion attacks. In the context of attacking AMC networks, the focus of this work, evasion attacks have been used to make slight intelligent changes to signals so that a trained AMC machine learning network misclassifies the signal [2, 6, 14, 16]. Therefore, these adversarial techniques can be used as a mitigation approach against eavesdroppers and malicious actors.

When developing these attacks, there is a natural tradeoff that arises between security and the intended application. For example, while the goal of an evasion attack against an AMC machine

learning algorithm is to cause a misclassification and/or reduce user confidence, it is important that the perturbed signal still accomplish its intended use of still being successfully received by its intended target. In the field of image recognition, this manifests in the idea that an image perturbed by an adversarial attack should still be easily discernible, and even viewed as untouched, by a human observing the image [8].

The adversarial scenario considered in this work is illustrated by Figure 1. As previously mentioned, balancing the two conflicting goals of evasion of an eavesdropper and successful communication by an intended receiver is difficult and has been examined in previous work [4, 5, 10, 15]. In this work, the idea of successful perception of the signal at the receiver is driven using metrics such as bit error rate (BER) that indicate the success of the communication. In addition, this work presents a novel form of perception to be considered alongside BER, that of spectral integrity. The previous works in this area have shown that adversarial perturbations naturally tend to manifest out of the main lobe of the signal and lead to adversarial signals that do not hold well the same spectral shape as the original signal [5]. This change in the spectral shape of the signal poses a problem to the success of the attack as the eavesdropper could leverage preprocessing stages to reduce the impact of the perturbation, such as with a filter, and potentially can lead to increased likelihood of detection that an attack is taking place since the spectral shape does not appear benign.

This work introduces a new loss metric for training machine learning based adversarial evasion attacks that helps maintain spectral integrity of the adversarially perturbed signal while still successfully achieving evasion and solid communication. Section II of this paper first provides a background on previous work done in this field and the particular evasion attack method used in this work. Section III introduces candidate spectral integrity loss metrics and provides relevant performance analysis. Finally, Section IV concludes this work and discusses future work.

2 BACKGROUND

Without proper care, evasion attacks used to fool an AMC machine learning algorithm generally have a drastic negative impact on the communication link between the transmitter and intended receiver. Recently, work by the authors and others have examined how these attacks can be improved in order to provide a better balance between these two conflicting goals. Hameed et. al. [10] accomplished this by introducing a gradient descent training method to craft signal perturbations that utilize a combined target function that considers both evasion performance and BER. While BER is non-differentiable, and thus not suited to gradient based learning approaches, a gradient is estimated using simultaneous perturbation stochastic approximation (SPSA). This approaches offers improvement over previous methods where the perturbation was simply power limited in the hope that this would lead to decreased BER. Flowers et al. improved upon these prior works through the development of a so-called "communications-aware" attack [5].

For the communications-aware attack, an Adversarial Residual Network (ARN) is leveraged in order to learn to make intelligent signal perturbations that balance the two opposing goals of evasion and communication. This approach utilizes three separate loss functions to accomplish this: adversarial loss, communication loss, and power loss. These three losses are each weighted and summed together to guide the ARNs learning process. The work of [4] expanded on the communications aware architecture first introduced in [5] in order to better utilize forward error correction (FEC), but it was found that the changes to the loss functions and transmitter architecture provided improvements beyond just utilizing FEC. The training framework presented in [4], and illustrated in Figure 2, serves as the foundation for the work presented in this paper.

As shown in Figure 2, the considered approach utilizes an Adversarial Mutation network (AMN) that is trained to create an intelligently perturbed signal given the original signal as input using a convolutional neural network (CNN). This adversarial signal is what is transmitted to the intended receiver and intercepted by the eavesdropper. It is assumed here that the eavesdropper utilizes an AMC network with the architecture described in [18] trained for BPSK, QPSK, 8PSK, 16QAM, and 64QAM. Each AMN is trained to create adversarial signals for just one modulation scheme at at a time. As previously mentioned, the AMN developed in [4] utilizes three loss functions to train the AMN network, namely:

- *Adversarial Loss*: prioritizes the AMN's ability to successfully learn to avoid classification by the eavesdropper. It is calculated using the confidence of the eavesdropper in the true source modulation, p_s, determined using the output of the final softmax layer in the eavesdropper's AMC.
- *Communication Loss*: prioritizes the AMN's ability to successfully learn to maintain the communication link between the transmitter and friendly receiver. It does this by using the calculated BER, b_r, as well as the error vector magnitude (EVM) between the clean symbols and the perturbed symbols, defined as $|S_{tx} - S_{tx+p}|$. BER is calculated using the original bits at the transmitter and the final bits decoded at the receiver after undergoing AWGN channel effects. The AWGN channel adds random noise between 0-20 dB. In this work it is assumed that the transmitter has access to the receiver in order to know the bits received. The BER is the true metric that the network wants to minimize, but is non-differentiable, so the EVM, which is differentiable, acts as a proxy for the BER and provides a gradient indicating the direction the weights should update. The BER then provides the magnitude of the update along this gradient.
- *Power Loss*: prioritizes the AMN's ability to learn to minimize the power of the perturbation so that the adversarial signal is close to the original signal. It does this by using the inverse of the signal-to-perturbation ratio (SPR).

During the AMN's training process, these three losses are each scaled and then summed together to create the total loss. These scaling factors allow for finer balancing between the communication and evasion goals. The scaling factors are α for adversarial loss, β for communications loss, and γ for power loss. More specifically, increasing a scaling factor relative to the others during training results in the corresponding loss being more highly prioritized. These loss constants are restricted such that they must sum to 1. Finally, the three loss functions are designed such that they all converge to zero. Therefore the network learns to minimize the loss values during training. Currently, these loss constants are estimated during

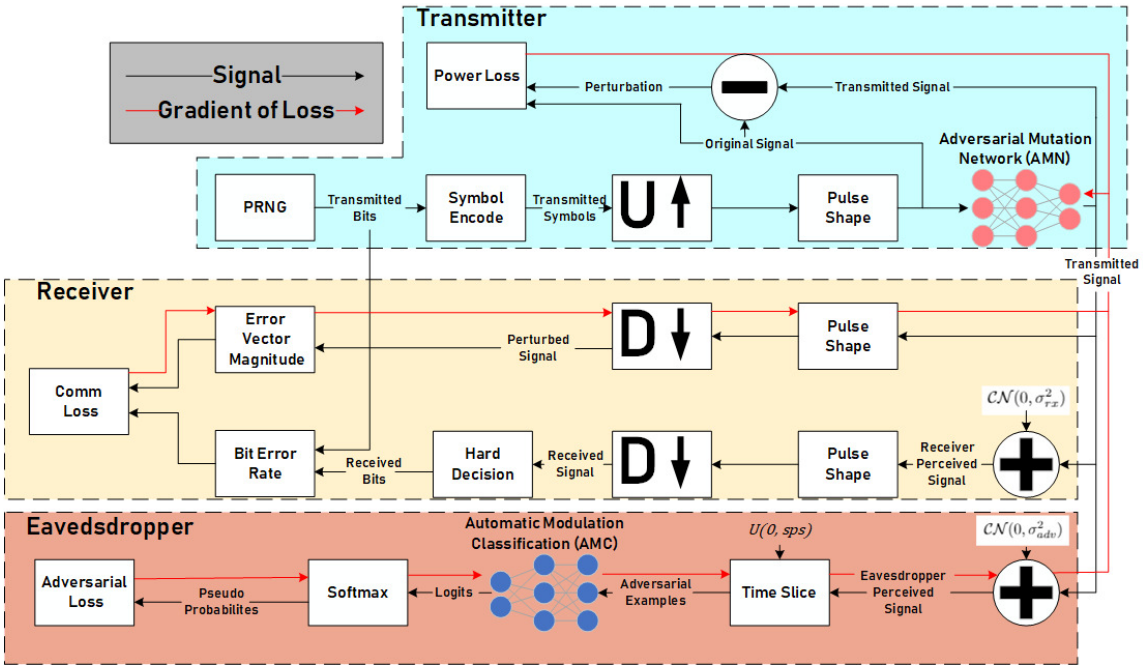

Figure 2: The training process that this work builds off of. It utilizes three losses, evasion loss, communication loss, and power loss each calculated at the eavesdropper, receiver, and transmitter respectively [4]. This work will replace the power loss calculated in the transmitter, leaving the others untouched.

training based on the rough needs of the system, such as whether evasion, communication, or power should be more important. A more exhaustive look into the best way to determine the values of the constants is left to future work. During the training process, the total loss is back-propagated through the CNN of the AMN to update the weights in order to create the most effective adversarial signal. The optimization technique used is Adam. To summarize, the loss functions are defined below:

$$\mathcal{L}_{\text{total}} = \alpha \mathcal{L}_{\text{adv}} + \beta \mathcal{L}_{\text{comm}} + \gamma \mathcal{L}_{\text{pwr}} \tag{1}$$

$$\mathcal{L}_{\text{adv}} = -log(1 - p_s) \tag{2}$$

$$\mathcal{L}_{\text{comm}} = b_r \times EVM(S_{tx}, S_{tx+p}) \tag{3}$$

$$\mathcal{L}_{\text{pwr}} = \frac{1}{E_s/E_p} = \frac{E_p}{E_s} \tag{4}$$

The architecture changes specified in [4], originally designed for use of forward error correction coding on the signal, allowed for improved spectral shape over the results seen in [5]. This improvement was predominantly due to improvements in the power loss metric and the usage of AMN as opposed to an ARN.

In this work, the same framework described above is used. However, here the power loss is replaced with a novel loss metric termed spectral deception loss. The goal of this loss will be to more explicitly train the network to create adversarial signals that follow the same spectral shape as the original signal, while still balancing between the conflicting goals of evasion and intended communications. The rest of the architecture, including the adversarial and communication loss, remains unchanged.

3 SPECTRAL DECEPTION LOSS

In this section, a variety of candidate spectrum deception loss metrics are presented, and their different impacts on the adversarial signal's spectral content are analyzed, along with its performance on eavesdropper evasion and intended communication capabilities. As previously discussed, it is desirable for the adversarial signal to have a similar spectral shape as the original signal so that it avoids detection and defensive capabilities. In this work, we determine this similarity through the power spectral density (PSD) and associated phase plot of the original signal, perturbation, and combined adversarial signal. Due to space considerations, only the PSD and not the phase plots are shown as the PSD provides a much better indication of success.

3.1 Examining the Necessity of Deception Loss

In the previous work, there was uncertainty over whether the power loss metric was sufficiently useful at providing the desired intent of maintaining the original shape of the signal. This was due partially to the fact that the two main performance metrics, BER and evasion classification success, were driven directly by the communication and adversarial loss, respectively, and not by the power loss. Additionally, the power loss and communication loss were shown to push the network to converge in the exact same way for these metrics, as is shown in Figure 3, leading to unnecessary redundancy among these two losses.

It makes sense that these two losses would provide similar results for the chosen performance metrics. However, observation of the PSD of the resulting adversarial signal when prioritizing each loss highlights the true differences between them. An example of

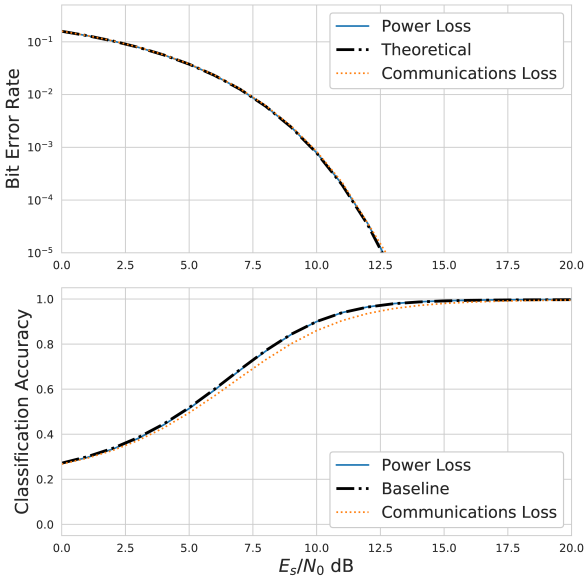

Figure 3: The BER and eavesdropper classification accuracy for QPSK adversarial signals when trained with either only communications loss or only power loss. The values are plotted over a range of 0-20 dB SNR. The theoretical values for the BER and classification accuracy of QPSK are shown.

this difference is shown in Figure 4. Prioritizing the power loss results in a PSD shape for the perturbation that is similar to the original signal, only less powerful. On the other hand, prioritizing the communication loss results in a PSD that is more jagged in the center lobe and has significant side lobe content. From this result, it can be observed that the power loss metric steers the training of the AMN to keep the spectral shape of the original signal while the communication loss metric disregards the original signal shape as long as the intended receiver is minimally impacted.

While the power loss appears to provide the exact behavior desired to maintain spectral integrity, this is only true under an ideal scenario. In the power loss result shown in Figure 4, the power loss is the only loss prioritized. However, when being balanced with the communication and evasion losses, the shape, while still an improvement on previous work, no longer resembles a clean signal and has some side lobe content [4]. Spectral deception loss is introduced as a solution to this problem so that the spectral integrity can be preserved even when successfully evading and communicating. The deception loss will operate in the frequency domain and thus allows for the AMN to better control the frequency content of the signal as opposed to the prior power loss metric that controls the time content of the signal. This should allow for better success in shaping the signal. As mentioned previously, this control over the spectral shape is desirable so that the attack can better avoid defenses such as filtering in the preprocessing stage of the eavesdropper. Previous work resulted in perturbations that had significant content in the side lobe. Such a perturbation could be weakened by a low pass filter that would remove this side lobe content and potentially render the attack ineffective. By forcing the

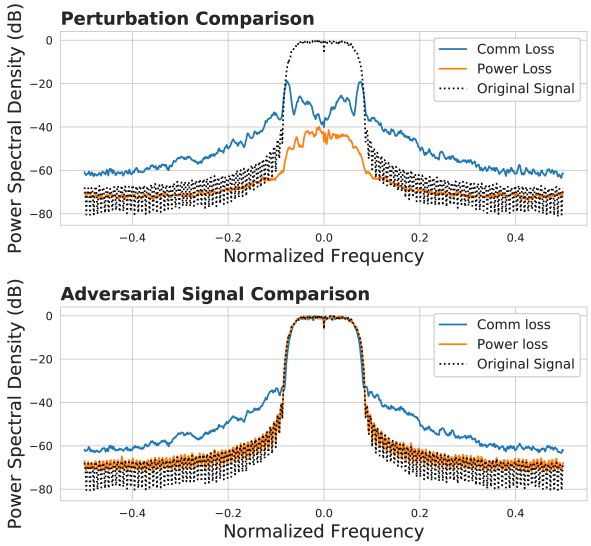

Figure 4: The PSD for both a perturbation created using only communication loss and one created using only power loss compared to the PSD of the original signal.

perturbation to be more in lobe, the deception loss should help the attack remain robust to these forms of filtering and defense.

3.2 Deception Loss

The deception loss method to be discussed within this work is based upon the frequency domain characteristics of the signal. More specifically, the proposed deception loss function operates on the Fast-Fourier Transform (FFT) of both the perturbation and original signal. This is done so that the perturbation lies more in-band and thus the adversarial signal will exhibit less side-lobe content and appear more benign. A function must be used in order to quantify the difference between the two FFTs. Two functions, Mean Squared Error (MSE) and Huber, are examined in this paper for their potential use in the deception loss.

3.2.1 MSE FFT Loss. MSE is a regression loss function that determines the difference between expected and actual values. In this paper, MSE is used as the average squared difference between the FFTs of the original signal and the perturbation. MSE is defined as:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{5}$$

where y is the value of the original signal and $\hat{y}$ is the value of the perturbation. After calculation, the loss was normalized such that $0 \leq \text{MSE} \leq 1$ to better align with the communication and adversarial loss values.

3.2.2 Huber FFT Loss. Although MSE is a good comparison metric for two functions, it is often heavily influenced by outliers. Huber loss mitigates the affect of outliers through an adjustable delta value, δ. If the absolute difference between the expected and actual value is less than δ, then Huber loss calculates their difference using an equation similar to MSE. Otherwise, the affect of the outlier is

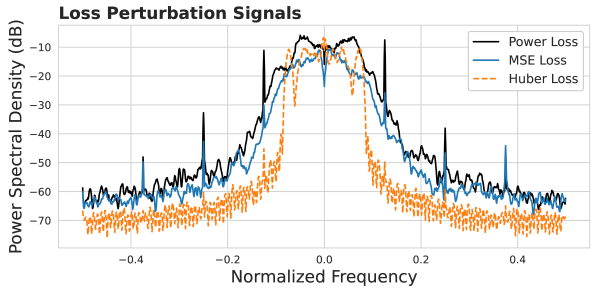

Figure 5: The PSD for the perturbations created by the original power loss and both the MSE and Huber loss methods on the FFT for BPSK signals.

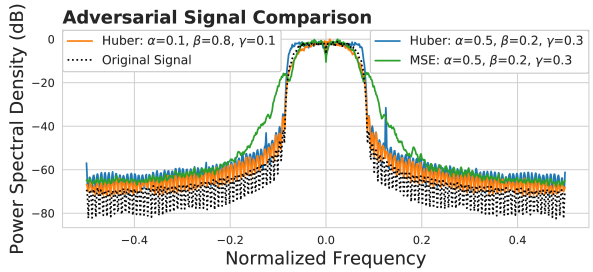

Figure 6: The PSD for the adversarial signals created by the MSE and Huber loss methods on the FFT for QPSK signals.

adjusted using the Mean Absolute Error (MAE) function. The Huber loss function is shown below.

$$\begin{cases} \frac{1}{2}(y-\hat{y})^2 & |y-\hat{y}| \leq \delta, \\ \delta|y-\hat{y}| - \frac{1}{2}\delta^2 & \text{otherwise} \end{cases} \tag{6}$$

where y is the value of the original signal and $\hat{y}$ is the value of the perturbed signal. Equation 6 specifies the function used to calculate the difference between two corresponding points in the FFTs of the original signal and perturbation. These differences are then summed and divided by the total number of points to obtain an average, like what is done with MSE. The value of δ used in this work is 1. Similar to MSE, Huber loss is normalized such that the loss value is contained between 0 and 1.

3.3 Results

The primary qualitative metric used when examining the success of the various spectral deception loss methods at maintaining the spectral shape was visual inspection of the PSD. Quantitative metrics used to validate the success of the considered metrics include the BER of the intended communications link and the achieved reduction in classification accuracy of the eavesdropper. The results presented in this section are predominantly examined with AMNs trained for QPSK modulated signals. However, other modulation schemes were also tested and exhibited the same characteristics. The eavesdropper's AMC used in this work was trained on BPSK, QPSK, 8PSK, 16QAM, and 64QAM.

As mentioned previously, this FFT-based approach was tested using both the MSE loss function and the Huber loss function. Figure 5 shows the resulting PSDs of just the perturbation for the MSE loss,

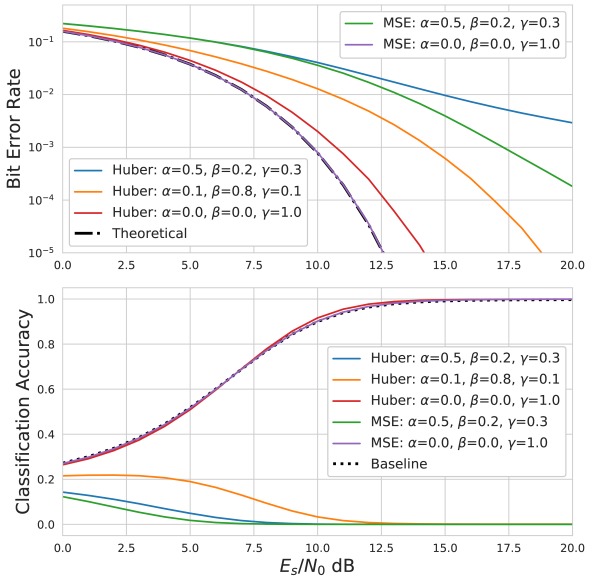

Figure 7: The BER and eavesdropper classification accuracy for QPSK adversarial signals when with the deception loss that is done on the FFT using both Huber and MSE loss. The signals correspond to those shown in Figure 6.

Huber loss, and the original power loss from the prior work. This figure illustrates that there is slight improvement with the MSE method over the power loss metric, but very minimal. However, the Huber loss method exhibits much better behavior over the power loss metric given that the shape of the perturbation is much more in-band to the original signal. This difference is due to the fact that the Huber loss is able to better handle situations of extreme error, which can occur during the training process especially at the start of training. Figure 6 shows the PSDs of the resulting adversarial signals. As can be seen from this figure, there is a trade off between the MSE method and the Huber method with respect to side lobe growth vs. main lobe corruption.

As expected, this trade-off in spectral shape performance comes at the detriment of intended communication performance. Figure 7 shows the BER and eavesdropper classification accuracy over the SNR range of 0-20 dB of the two methods, along with the theoretical QPSK bit error rate with no perturbation added. When using loss constants of $\alpha = 0.5$, $\beta = 0.2$, and $\gamma = 0.3$, the BER rate for the Huber method is much worse than that of the MSE method. Additionally, when the MSE deception loss is the only loss considered during training (i.e. α and β are set to 0), the BER converges to the theoretical curve, which does not occur for the Huber loss. However, by adjusting the loss constants, the communication performance can be made better as is shown by the Huber result with $\alpha = 0.1$, $\beta = 0.8$, and $\gamma = 0.1$. Naturally, this does lead to worse evasion performance. Interestingly, in Figure 6 it can be observed that the resulting spectral shape of the adversarial signal does not seem to drastically change for this second Huber trial even though the deception loss is less prioritized. This shows that the constants can be adjusted to meet the needs of the attack and that when using

Huber loss in the deception loss, the spectral shape can be maintained even when less prioritized (so more priority can be spent on evasion or communication improvement).

4 CONCLUSION AND FUTURE WORK

The results of this work show the benefit of utilizing a spectral deception loss metric within the considered machine learning based adversarial evasion attack. The considered FFT-based methods of developing this metric provided solid improvements over the prior work and can be used as the foundation for future work. Utilizing the FFT, two loss metrics were considered, namely MSE and Huber losses. Performance analysis demonstrated that the Huber loss was more successful at maintaining the spectral shape of the original signal over the MSE loss, at the cost of decreased intended communication performance.

While these results show promise, there is still much future work to develop the concept further. The various deception loss methods presented in this work are intended as starting points and improving upon these may offer greater success. For example, one simple adaptation could come in the form of completing a more exhaustive parameter search over the configurations for the deception loss. Additionally, other functions than the FFT investigated here could be used to determine and quantify the difference between the original signal and the adversarial signal. For example, minimizing the difference between the resulting PSDs and associated phases could be examined. Finally, while mean squared error (MSE) and Huber are good for determining the difference between corresponding elements in an array of data, such as with time domain samples, they may not be the most appropriate for the frequency domain. Other functions, such as Fréchet distance, may provide better comparisons of similarity and should be further studied.

The predominant method used in this work to determine success of the loss was to qualitatively observe if the perturbation was concentrated in the main lobe of the signal. While this may be sufficient in determining whether a human operator can detect the adversarial signal, future work should examine whether this adapted attack framework would be effective in evading detection by a machine learning algorithm aimed at detecting these attacks. Additionally, previous work has assumed oversampling of the signal by the eavesdropper which provides a larger attack vector for the evasion attack in terms of available bandwidth outside of the signal's main lobe. Future work should loosen this assumption in order to better test the success of the deception loss. Recent work has focused on strategies that make the classifier networks more robust against attacks such as utilizing curriculum training [9]. Future work should examine the success of evasion attacks against such defensive techniques when employing the deception loss.

While the concept of a spectral deception loss is an extremely new area of focus, this work has shown that it is one that offers great potential in the effort to mask the limitations and distinguishing characteristics of existing evasion attacks.

5 ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant Number 1303297. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This material was also aided as part of an undergraduate research program sponsored by the Naval Engineering Education Consortium and NSWC Crane.

REFERENCES

[1] Nof Abuzainab, Tugba Erpek, Kemal Davaslioglu, Yalin E. Sagduyu, Yi Shi, Sharon J. Mackey, Mitesh Patel, Frank Panettieri, Muhammad A. Qureshi, Volkan Isler, and Aylin Yener. 2019. QoS and Jamming-Aware Wireless Networking Using Deep Reinforcement Learning. *IEEE Military Commun. Conf. (MILCOM)* (2019).

[2] Samuel Bair, Matthew DelVecchio, Bryse Flowers, Alan J Michaels, and William Chris Headley. 2019. On the limitations of targeted adversarial evasion attacks against deep learning enabled modulation recognition. *ACM Workshop on Wireless Security and Machine Learning (WiseML 2019)* (2019).

[3] Kemal Davaslioglu and Yalin E. Sagduyu. 2019. Trojan Attacks on Wireless Signal Classification with Adversarial Machine Learning. *IEEE Int. Symp. on Dynamic Spectrum Access Networks (DySPAN)* (2019).

[4] Matthew DelVecchio, Bryse Flowers, and William C. Headley. 2020. Effects of Forward Error Correction on Communications Aware Evasion Attacks. *In Review* (2020).

[5] Bryse Flowers, R Michael Buehrer, and William C Headley. 2019. Communications Aware Adversarial Residual Networks for Over the Air Evasion Attacks. *IEEE Military Commun. Conf. (MILCOM)* (2019).

[6] Bryse Flowers, R Michael Buehrer, and William C Headley. 2019. Evaluating Adversarial Evasion Attacks in the Context of Wireless Communications. *IEEE Trans. on Info. Forensics and Security* (2019).

[7] Selen Gecgel, Caner Goztepe, and Gunes Karabulut Kurt. 2019. Jammer Detection Based on Artificial Neural Networks: A Measurement Study. *ACM Workshop on Wireless Security and Machine Learning (WiseML 2019)* (2019).

[8] Ian Goodfellow, Jonathon Shelens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. *Int. Conf. on Learning Representations* (2015).

[9] Muhammad Zaid Hameed, Andras Gyorgy, and Deniz Gunduz. 2019. The Best Defense Is a Good Offense: Adversarial Attacks to Avoid Modulation Detection. arXiv:arXiv preprint arXiv:1902.10674

[10] Muhammad Zaid Hameed, Andras Gyorgy, and Deniz Gunduz. 2019. Communication without Interception: Defense against Modulation Detection. *IEEE Global Conf. on Signal and Info. Processing (GlobalSIP)* (2019).

[11] Alharbi Hazza, Mobien Shoaib, Saleh A. Alshebeili, and Alturki Fahad. 2013. An overview of feature-based methods for digital modulation classification. *1st Int. Conf. on Commun., Signal Processing, and their Applications (ICCSPA)* (2013).

[12] William C. Headley, Jesse D. Reed, and Claudio R. C. M. da Silva. 2008. Distributed Cyclic Spectrum Feature-Based Modulation Classification. *IEEE Wireless Commun. and Netw. Conf.* (2008).

[13] Ling Huang, Anthony D. Joseph, Blaine Nelson, Benjamin I.P. Rubinstein, and J.D. Tygar. 2011. Adversarial Machine Learning. *ACM Workshop on Security and Artificial Intelligence* (2011), 43 – 58.

[14] Brian Kim, Yalin E. Sagduyu, Kemal Davaslioglu, Tugba Erpek, and Sennur Ulukus. 2020. Channel-Aware Adversarial Attacks Against Deep Learning-Based Wireless Signal Classifiers. *arXiv preprint arXiv:2005.05321* (2020).

[15] Brian Kim, Yalin E. Sagduyu, Kemal Davaslioglu, Tugba Erpek, and Sennur Ulukus. 2020. How to Make 5G Communications "Invisible": Adversarial Machine Learning for Wireless Privacy. *arXiv preprint arXiv:2005.07675* (2020).

[16] Brian Kim, Yalin E. Sagduyu, Kemal Davaslioglu, Tugba Erpek, and Sennur Ulukus. 2020. Over-the-Air Adversarial Attacks on Deep Learning Based Modulation Classifier over Wireless Channels. *Conf. on Info. Sciences and Systems (CISS)* (2020).

[17] Zachary Langford, Logan Eisenbeiser, and Matthew Vondal. 2019. Robust Signal Classification Using Siamese Networks. *ACM Workshop on Wireless Security and Machine Learning (WiseML 2019)* (2019).

[18] Timothy J. O'Shea, Johnathan Corgan, and T. Charles Clancy. 2016. Convolutional Radio Modulation Recognition Networks. *Commun. in Computer and Info. Science* 629 (2016).

[19] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. 2016. The limitations of deep learning in adversarial settings. In *IEEE European Symp. on Security and Privacy (EuroS&P)*.

[20] Yi Shi, Tugba Erpek, Yalin E. Sagduyu, and Jason H. Li. 2018. Spectrum Data Poisoning with Adversarial Deep Learning. *IEEE Military Commun. Conf. (MILCOM)* (2018).

[21] Nathan E. West and Tim J. O'Shea. 2017. Deep architectures for modulation recognition. *IEEE Int. Symp. on Dynamic Spectrum Access Networks (DySPAN)* (2017).

[22] Jack L. Ziegler, Robert T. Arn, and William Chambers. 2017. Modulation recognition with GNU radio, keras, and HackRF. *IEEE Int. Symp. on Dynamic Spectrum Access Networks (DySPAN)* (2017).