WIDEFT: A CORPUS OF RADIO FREQUENCY SIGNALS FOR WIRELESS DEVICE FINGERPRINT RESEARCH

Abu Bucker Siddik[1], *Dawson Drake*[1,2], *Thomas Wilkinson*[1,2], *Phillip L. De Leon*[1], *Steven Sandoval*[1], and *Margaret Campos*[2]

New Mexico State University
[1]Klipsch School of Electrical and Computer Engineering
[2]Physical Science Laboratory
Las Cruces, New Mexico, U.S.A.
{siddik,dvd82000,wthomasw,pdeleon,spsandov,mcampos}@nmsu.edu

ABSTRACT

Wireless network security may be improved by identifying networked devices via traits that are tied to hardware differences, typically related to unique variations introduced in the manufacturing process. One way these variations manifest is through unique transient events when a radio transmitter is activated or deactivated. Features extracted from these signal bursts have in some cases, shown to provide a unique "fingerprint" for a wireless device. However, only recently have researchers made such data available for research and comparison. Herein, we describe a publicly-available corpus of radio frequency signals that can be used for wireless device fingerprint research. The WIDEFT corpus contains signal bursts from 138 unique devices (100 bursts per device), including Bluetooth- and WiFi-enabled devices, from 79 unique models. Additionally, to demonstrate the utility of the WIDEFT corpus, we provide four baseline evaluations using a minimal subset of previously-proposed features and a simple ensemble classifier.

Index Terms— Wireless networks, Network security, Signal analysis

1. INTRODUCTION

Providing security in a wireless network environment is a challenging problem because access to network resources is possible without being physically connected. One way to enhance security is through unique identifiers associated to particular network devices without considering easily forgeable identifiers such as MAC addresses and user credentials [1]. Prior research has demonstrated the ability to identify a device via physical, hardware-level traits due to variations/imperfections in the circuitry associated with manufacturing process [2]. One way these variations/imperfections manifest is through unique transient events when an wireless transmitter is activated or deactivated. Features extracted from these transient events may provide a unique "fingerprint" for a wireless device that is similar to a biometric for human identification [3]. Thus, device fingerprinting can offer an additional defensive layer of security beyond conventional (software) protocols for authentication [4, 5].

More specifically, hardware-based device-level fingerprinting offers additional security and safeguards against attacks such as device cloning, message replay, and spoofing attacks [6]. On the other hand, fingerprinting may also be used offensively to gain information about network operations or to gain information about specific network users [6]. A skilled attacker, for example, may be able to exploit a device fingerprint to violate sender anonymity, in order to associate a transmission to a specific sender. For example, FM transmitters have been shown in many cases, to have relatively short transient characteristics directly following when the transmitter is activated or "keyed" [7]. Potentially, these transient characteristics are similar enough that the device model may be identifiable, and in the best case are unique enough to allow identification of an individual device. Recently, the widespread availability/use of software defined radios (SDRs) has lowered the barrier to adversarial cloning and spoofing [8]. One potential defense against these attacks could be device

fingerprinting.

In this work, we describe a publicly-available data corpus for WIreless DEvice FingerprinTing (WIDEFT) research. In addition, we provide details of a feature vector and classifier in order to provide baseline evaluation results using the WIDEFT corpus. In Section 2, we provide some background information on an existing corpus and our motivation for WIDEFT. This section carefully describes the devices for which we collect signal bursts from, the data collection environment, and our procedure for collecting the signal data. In Section 3, we describe the signal postprocessing which includes segmentation and the organization of the resulting files. In Section 4, we provide results from four baseline evaluations using a minimal subset of previously-proposed features and a simple ensemble classifier. In Section 5, we provide a short discussion on the WIDEFT corpus and the evaluation results and finally in Section 6, we conclude the article.

2. WIDEFT CORPUS

To the best of our knowledge there exists at least one recent paper which describes a publicly-available data corpus for wireless device fingerprinting. In [9], the authors focus their data collection on Bluetooth devices which consisted of 86 individual smartphones from a set of 27 different models. The data consists of wireless signals acquired in a lab environment using a network analyzer at different sample rates. The signals are subsequently processed by transformation to the analytic signal via the Hilbert transform (HT), digitally downconverted, and segmented to preserve the ON transient and a portion of the steady-state which immediately follows. The authors also present experimental results for classification of the devices using features extracted from the signal transient [9].

Unlike the data set in [9], WIDEFT includes not only Bluetooth devices (e.g. wireless ear buds, keyboards, mice) but also WiFi devices (e.g. wireless routers, wireless media servers) and other 900 MHz devices (e.g. wireless clickers). In some devices, such as laptops and smartphones, both Bluetooth and WiFi signals were acquired. Unlike [9], we collect a complete radio burst consisting of the ON transient, steady-state portion, and OFF transient as all three portions may offer discriminating features to allow for improved fingerprinting. Our motivation for collecting a wide variety of devices is to allow research into the uniqueness (or lack thereof) of device fingerprints. This corpus of wireless bursts is made publicly-available to researchers on Zenodo (DOI 10.5281/zenodo.4110980) [10].

2.1. Device Types and Distribution

The WIDEFT data corpus is broken up into three wireless types: Bluetooth, WiFi, and Other. The data is collected from 138 devices spanning 79 models[1] and is summarized in Table 1. Each of the 138 devices is assigned a unique device identifier. Table 2 provides information on which models have more than one device represented in the WIDEFT corpus[2]. Because some models/devices are capable of more than one wireless type and/or radio modes, 147 data captures were made. Each capture contains a set of 100 bursts. An index file [10] accompanies the WIDEFT corpus and provides additional device details such as make/manufacturer, model, center frequency, etc.

Table 1. WIDEFT Wireless Type Distribution. For additional device information such as make/manufacturer, model, center frequency, etc the reader is referred to the accompanying index file in the corpus.

Wireless Type	Num. Unique Models	Num. Unique Device IDs
Bluetooth (2.4 GHz)	64	81
WiFi (2.4 GHz, 5 GHz)	49	55
Other (900 MHz)	1	2
Total[1]	**79**	**138**

2.2. Data Collection Environment

Our data acquisition system consisted of the USRP B210 SDR [11], a laptop computer, an antenna at the operating frequency of the device, and the device being examined. The laptop and SDR were connected via USB with the antenna connected to the SDR via a coax cable connected to the RX port. A stand was often used to hold the antenna steady to prevent any movement in the antenna during collection and to keep the collections as consistent as possible. To achieve the best SNR possible the bursts were collected inside an anechoic chamber located on the New Mexico State University

[1]Some models have both Bluetooth and WiFi wireless types hence why the number of unique Bluetooth models and unique WiFi models do not sum to the total number of unique models.

[2]In some cases of earbud pairs, the left earbud has a different model number (e.g. A2031) than the right earbud (e.g. A2032). For our purposes, we consider these two devices as having the same model (e.g. A2031/A2032).

Table 2. Models in WIDEFT where data is collected from more than one device. Models in WIDEFT not listed below have data collected from only one device.

Wireless Type	Mfr. Model	Num. Devices
Bluetooth (2.4 GHz)	Apple A2031/A2032	8
	Aimson Q88	3
	Apple A1367	3
	Apple A2084/A2083	2
	Beats A2047/A2048	2
	Dell PP32LB	2
	Haylou GT1	2
	Lenovo 81QB	2
	MPOW BH437A	2
WiFi (2.4 / 5 GHz)	Aimson Q88	3
	Apple A1367	3
	Dell PP32LB	2
	Lenovo 81QB	2
Other	iClicker RLR14	2

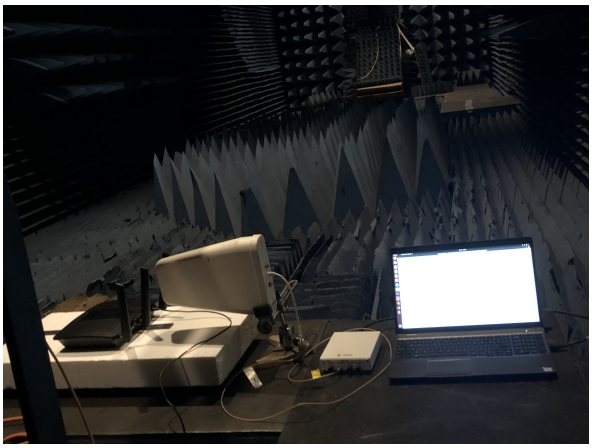

Fig. 1. Photograph of the data acquisition setup consisting of a software defined radio (SDR) model USRP B210 connected to a PC using GNU Radio for radio device control and data acquisition.

campus. Our collection environment and data acquisition system, shown in Fig. 1, is similar to that used in [9]. For each device, we determined an orientation of the antenna which resulted in the highest signal power. A sampling rate of 56 MHz was used by the data acquisition system and the raw data recording was saved on a 3 GB RAM disk to ensure that the data could be saved at the fastest speed possible.

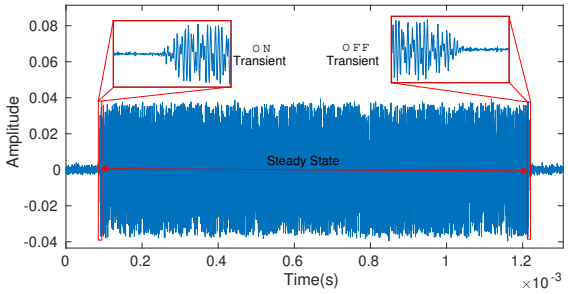

Fig. 2. Plot of the signal burst with ON on and OFF transients zoomed in for more detail. This figure captures the period before (prefix) and after (suffix) the signal burst, the ON and OFF transients, and the steady-state portion of the signal.

3. POSTPROCESSING AND DATA ORGANIZATION

3.1. Signal Post-Processing

In order to segment bursts from the raw recordings, we developed a MATLAB script to automate the burst detection. This script applied changepoint detection to the power spectrum to determine the beginning and end of the each radio burst. The signal segments provided in WIDEFT include 5000 samples before and after each burst to ensure the entire burst is captured. Fig. 2 shows an example signal segment with the ON/OFF transients and steady state, as well as 5000 samples before and after the transients. The signal bursts are saved as 16 bit I-Q sample pairs in `.sc16` file format. This process is repeated for each device until 100 distinct bursts have been collected and processed.

3.2. Data Organization

The data files in WIDEFT are organized into a directory tree with the following path:
`WIDEFT/<WIRELESS_TYPE>/<FREQUENCY>/<PROTOCOL>/<DEVICE_MAKE>/<DEVICE_MODEL>_<DEVICE_NUMBER>/<BURST_ID>.sc16` where `WIRELESS_TYPE` is 'Bluetooth', 'WiFi', or 'Other'; `FREQUENCY` (WiFi only) is '2.4GHz' or '5GHz'; `PROTOCOL` (WiFi only) is '802.11a', '802.11b', '802.11g', '802.11n', '802.11ac', or 'Auto'; `DEVICE_MAKE` is the manufacturer of the device; `DEVICE_MODEL` is the model number associated with the device; `DEVICE_NUMBER` is used to differentiate multiple devices with the same model family; and `BURST_ID` is an integer from 1 to 100.

The following are some examples of file paths and names for a single file.

Bluetooth example:
`WIDEFT/Bluetooth/Apple/A1296_1/1.sc16` translates to `Bluetooth` wireless type, manufactured by `Apple`, model number `A1296`, device `1`, burst ID `1`, `sc16` file.

WiFi example:
`WIDEFT/WiFi/2.4GHz/Auto/Dell/M4800_1/15.sc16` translates to `WiFi` wireless type, `2.4GHz` frequency, `Auto` protocol, manufactured by `Dell`, model number `M4800`, device `1`, burst ID `15`, `sc16` file.

Other example:
`WIDEFT/Other/iClicker/RLR14_2/50.sc16` translates to `Other` wireless type, manufactured by `iClicker`, model number `RLR14`, device `2`, burst ID `50`, `sc16` file.

The `index.csv` file contains the associated information about the devices found in the corpus. N/A is used in the wireless protocol column for Bluetooth and Other devices. A '*' is used as a placeholder for the different burst number in the file name and path. The index file is broken up into the following columns in the given order: `DEVICE_ID`, `PATH`, `WIRELESS_TYPE`, `CENTER_FREQUENCY`, `PROTOCOL`, `DEVICE_MAKE`, `DEVICE_MODEL`, `FCC_ID`, `IC`, and `NOTES`.

4. CLASSIFICATION OF DEVICE FINGERPRINTS

In this section, we provide a baseline for classifier results using the the WIDEFT corpus. This includes a description of the feature vector whose elements are statistics of the instantaneous amplitude (IA), instantaneous phase (IP), and instantaneous frequency (IF) sequences during the ON transient. In addition, we describe the classifier and the results of the evaluations.

We consider four different evaluations. First, we consider two device identification evaluations: 1) device identification using all 138 devices[3] and 2) device identification using all eight available devices from model Apple A2031/A2032. Second, we consider two model identification evaluations: 3) model identification using all available devices per model and 4) model identification using only one device per model. For the remainder of the paper, we will denote these four evaluations as Evaluation 1, $\cdots$, Evaluation 4.

[3]For WiFi devices which have multiple operating modes, we use Auto and 2.4 GHz.

4.1. Feature Extraction

Previous research has considered features based on signal amplitude statistics [6], higher order statistics for use with common digital modulation schemes [12], cycle-frequency domain profile (statistical) for use with orthogonal frequency division multiplexing signals and tested on an IEEE 802.11a/g WLAN device [13], and normalized permutation entropy [14]. Other work uses time-frequency and time-scale methods. For example features are considered based on wavelet coefficients extracted from the transients [15], empirical mode decomposition and Haar wavelet decompositions [16], Hilbert-Huang Transform [17–19], ambiguity function and Wigner distribution [20], and the intrinsic time-scale decomposition [21].

For each signal burst we detect the ON transient and segment the burst as follows. The analysis segment begins at the start of the ON transient and consists of 100 samples. This is illustrated in the black frame in Fig. 3. For each segment we compute the corresponding IA, IP, and IF sequences [22]. We compute statistics i.e. mean, variance, skewness, and kurtosis from each instantaneous sequence (IA, IP, IF) and these form the elements of the 12-D feature vector [23, 24].

4.2. Classifier

For each baseline evaluation, the data was randomly partitioned into two subsets, 75% for training and 25% for testing. We use an ensemble of 1000 tree classifiers with bagging, implemented using the `TreeBagger` function in MATLAB.

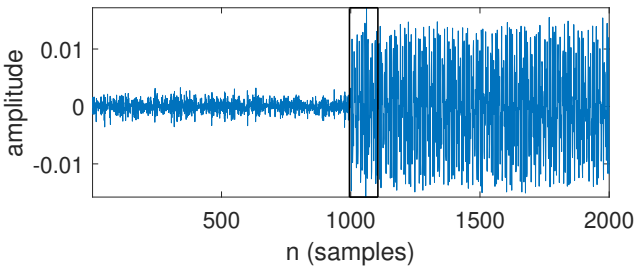

Fig. 3. An example of a single ON transient. The feature vector is computed using the region of the segment within the black frame.

4.3. Evaluations and Results

Table 3 shows the classification results for the four evaluation cases described earlier in this section. Each of

these results provides an initial or baseline classification accuracy for the WIDEFT corpus for these evaluations. We have chosen a minimal subset of previously-proposed features and a simple classifier for these baseline evaluations [23,24]. For Evaluation 1, where device identification was performed on all 138 devices, classification accuracy is 54.3%. For Evaluation 2, where device identification was performed on eight Apple A2031/A2032 devices, classification accuracy is 82.0% and the confusion matrix is given in Table 4. For Evaluation 3, where model identification was performed using all available devices per model, classification accuracy is 56.3%. For Evaluation 4, model identification using only one device per model, classification accuracy is 62.7%.

Evaluation	Accuracy (%)	Description
1	54.3	Device ID (all devices)
2	82.0	Device ID (Apple A2031/A2032)
3	56.3	Model ID (all devices)
4	62.7	Model ID (one device per model)

Table 3. Classification results for the four evaluations described in Section 4.

	Predicted							
Actual	*A2031_1*	*A2031_2*	*A2031_3*	*A2031_4*	*A2032_1*	*A2032_2*	*A2032_3*	*A2032_4*
A2031_1	**20**	1	·	·	·	4	·	·
A2031_2	1	**17**	·	·	4	·	3	·
A2031_3	·	·	**25**	·	·	·	·	·
A2031_4	·	·	·	**24**	·	·	·	1
A2032_1	1	4	·	·	**19**	·	1	·
A2032_2	8	·	·	1	·	**16**	·	·
A2032_3	1	4	·	·	1	1	**18**	·
A2032_4	·	·	·	·	·	·	·	**25**

Table 4. Confusion matrix for Evaluation 2 device identification (Apple A2031/A2032) described in Section 4.

5. DISCUSSION

In this work, we have introduced the WIDEFT corpus with the intent of advancing research in the area of wireless device fingerprinting and specific emitter identification. The corpus consists of signal captures from a 138 wireless devices with each capture consisting of 100 wireless bursts. Unlike the recent corpus developed in [9] which includes only the ON transient and a portion of the steady state, the WIDEFT corpus consists of entire bursts, including the ON transient, steady state, and OFF transient. The complete burst may allow for innovations in feature vector or classifier design for this research area. Additionally, while the corpus in [9] contains only Bluetooth smartphone devices, the WIDEFT corpus includes both Bluetooth and WiFi devices including smartphones, wireless speakers, headphones, earbuds, routers, keyboards, mice, and laptops. Finally, while the corpus in [9] contains signal captures from 27 unique smartphone models, the WIDEFT corpus contains signal captures from 79 unique models, including not just smartphones but other Bluetooth- / WiFi-enabled devices. This greater variety of device types and models allows a larger sample size when performing classifier evaluations and thus greater statistical power. On the other hand, increasing the number and variety of devices makes the problem more representative of the true "population" of wireless devices in current use.

Evaluation 1 shows that a feature vector formed from statistics of the IA, IP, and IF sequences during the ON transient can be used to identify more than half (54.3%) of the 138 unique devices. This experiment suggests that the fingerprint does have discriminating traits. Evaluation 2 considered a particular model, Apple A2031/A2032 which has the most example devices (eight) and shows that we can identify with 82.0% accuracy. This experiment suggests that the fingerprint can identify devices within the same model family. An inspection of the confusion matrix in Table 4 reveals that the false positives and false negatives for the first two devices (A2031_1, A2031_2) account for the majority of the errors. Evaluation 3 considered model identification using all devices. Unfortunately, because some models consist of several example devices and other models have only one example device, it is difficult to draw further conclusions. On the other hand, Evaluation 4 considered a model identification using a single device per model and thus has a balanced dataset (as compared to Evaluation 3). Assuming that in the feature space, the inter-model distance is much less than the intra-model distance, this can account for the increase in accuracy as compared to Evaluation 1. In all four evaluations, accuracy is far better than random guessing, however, there is considerable room for improvement in identification accuracy. The results of all four evaluations are in general agreement with findings in the existing literature.

6. CONCLUSIONS

In this paper, we have introduced the publicly-available WIDEFT data corpus for advancing research in the area of wireless device fingerprinting and specific emitter identification. The WIDEFT corpus contains signal bursts from 138 unique devices (100 bursts per device), including 900 MHz, Bluetooth, and WiFi devices from 79 unique models. Additionally, to give an initial baseline of classifier performance for both device identification and model identification tasks using the WIDEFT corpus, we performed four evaluations. In these evaluations, we used previously-proposed features based on statistics computed from the instantaneous amplitude (IA), instantaneous phase (IP), and instantaneous frequency (IF) sequences during the ON transient of the radio burst. These evaluations provide baseline results for future research in which more sophisticated features and classification schemes may be compared against.

7. REFERENCES

[1] Y. Zhang, W. Lee, and Y.-A. Huang, "Intrusion Detection Techniques for Mobile Wireless Networks," *Wireless Networks*, vol. 9, no. 5, pp. 545–556, Sep. 2003.

[2] S. Banerjee and V. Brik, *Encyclopedia of Cryptography and Security*. Boston, MA: Springer US, 2011, pp. 1388–1390.

[3] J. Hall, M. Barbeau, and E. Kranakis, "Detection of Transient in Radio Frequency Fingerprinting using Signal Phase," *Wireless and Optical Commun.*, pp. 13–18, 2003.

[4] B. Danev, D. Zanetti, and S. Capkun, "On Physical-Layer Identification of Wireless Devices," *ACM Computing Surveys*, vol. 45, no. 1, p. 6, 2012.

[5] Sieka and Bartlomiej, "Using Radio Device Fingerprinting for the Detection of Impersonation and Sybil Attacks in Wireless Networks," *Lecture Notes in Computer Science*, vol. 4357, p. 179, 2006.

[6] K. B. Rasmussen and S. Capkun, "Implications of Radio Fingerprinting on the Security of Sensor Networks," in *Proc. Int. Conf. Security and Privacy in Commun. Networks*, 2007, pp. 331–340.

[7] K. Ellis and N. Serinken, "Characteristics of Radio Transmitter Fingerprints," *Radio Science*, vol. 36, no. 4, pp. 585–597, 2001.

[8] T. D. Vo-Huu, T. D. Vo-Huu, and G. Noubir, "Fingerprinting Wi-Fi Devices using Software Defined Radios," in *Proc. ACM Conf. on Security and Privacy in Wireless and Mobile Networks*, 2016, pp. 3–14.

[9] E. Uzundurukan, Y. Dalveren, and A. Kara, "A Database for the Radio Frequency Fingerprinting of Bluetooth Devices," *Data*, vol. 5, Jun. 2020.

[10] (2020). [Online]. Available: https://zenodo.org/record/4110980

[11] (2020). [Online]. Available: https://files.ettus.com/manual/page_usrp_b200.html

[12] Y.-J. Yuan, Z. Huang, and Z.-C. Sha, "Specific Emitter Identification based on Transient Energy Trajectory," *Progress in Electromagnetics Research*, vol. 44, pp. 67–82, 2013.

[13] K. Kim, C. M. Spooner, I. Akbar, and J. H. Reed, "Specific Emitter Identification for Cognitive Radio with Application to IEEE 802.11," in *IEEE Global Telecommun. Conf.*, 2008, pp. 1–5.

[14] G. Huang, Y. Yuan, X. Wang, and Z. Huang, "Specific Emitter Identification Based on Nonlinear Dynamical Characteristics," *Canadian J. Elect. and Comp. Eng.*, vol. 39, no. 1, pp. 34–41, 2016.

[15] R. Hippenstiel and Y. Payal, "Wavelet Based Transmitter Identification," in *IEEE Sym. Sig. Process. and its Applications*, vol. 2, 1996, pp. 740–742.

[16] C. Song, J. Xu, and Y. Zhan, "A Method for Specific Emitter Identification based on Empirical Mode Decomposition," in *IEEE Int. Conf. Wireless Commun., Networking and Info. Security*, 2010, pp. 54–57.

[17] Y. Yuan, Z. Huang, H. Wu, and X. Wang, "Specific Emitter Identification based on Hilbert-Huang Transform-Based Time-Frequency-Energy Distribution Features," *IET Commun.*, vol. 8, no. 13, pp. 2404–2412, 2014.

[18] J. Zhang, F. Wang, Z. Zhong, and O. Dobre, "Novel Hilbert Spectrum-Based Specific Emitter

Identification for Single-Hop and Relaying Scenarios," in *IEEE Global Commun. Conf.*, 2015, pp. 1–6.

[19] J. Zhang, F. Wang, O. A. Dobre, and Z. Zhong, "Specific Emitter Identification via Hilbert-Huang Transform in Single-Hop and Relaying Scenarios," *IEEE Trans. Info. Forensics and Security*, vol. 11, no. 6, pp. 1192–1205, 2016.

[20] L. Li, H.-B. Ji, and L. Jiang, "Quadratic Time–Frequency Analysis and Sequential Recognition for Specific Emitter Identification," *IET Sig. Process.*, vol. 5, no. 6, pp. 568–574, 2011.

[21] C. Song, Y. Zhan, and L. Guo, "Specific Emitter Identification Based on Intrinsic Time-Scale Decomposition," in *IEEE Int. Conf. Wireless Commun. Networking and Mobile Computing*, 2010, pp. 1–4.

[22] S. Sandoval and P. L. De Leon, "The Instantaneous Spectrum: A General Framework for Time-Frequency Analysis," *IEEE Transactions on Signal Processing*, vol. 66, no. 21, pp. 5679–5693, 2018.

[23] X. Wang, Y. Zhang, H. Zhang, X. Wei, and G. Wang, "Identification and Authentication for Wireless Transmission Security Based on RF-DNA Fingerprint," *EURASIP Journal on Wireless Communications and Networking*, vol. 2019, no. 1, p. 230, 2019.

[24] A. Aghnaiya, A. M. Ali, and A. Kara, "Variational Mode Decomposition-Based Radio Frequency Fingerprinting of Bluetooth Devices," *IEEE Access*, vol. 7, pp. 144 054–144 058, 2019.

Biography of Presenting Author

Abu Bucker Siddik is a PhD student at the Klipsch School of Electrical and Computer Engineering at New Mexico State University (NMSU). He received his BS degree in electrical and electronic engineering from Khulna University of Engineering and Technology (KUET), Bangladesh in 2015. His research interests include signal processing, time-frequency analysis, atmospheric turbulence, and machine learning.