Applying the majority voting rule in acoustic detection and classification of drones

Rigel P. Fernandes[1], José A. Apolinário Jr.[1], Antonio L. L. Ramos[2], and José M. de Seixas[3]

***Abstract*— This paper discusses an approach to target detection and classification based on acoustic signals collected using one single microphone. This study has applications to sonar or any other sound event classification system. We divide the problem into two parts, namely feature extraction and target detection and classification. We use an optimization step based on human auditory uncertainty. We employ a majority voting rule for every set of feature vectors, i.e., an estimate is only performed if the majority agrees. We conducted experiments using a single channel of the AIRA-UAS dataset, a public database of raw drone noises collected with an array of microphones mounted on a drone. This dataset comprises many different kinematics, with different spectra. The features we used are based on the Mel-Frequency Cepstral Coefficients (MFCC) and the Short-Time Fourier Transform of raw signals. We used the K-Nearest Neighbors algorithm for classification and adopted the cross-validation strategy to evaluate the method. We observed that the use of MFCC results in less biased estimations, which favors the voting strategy. The detection in the proposed method reached a probability of false positive near 0%, even with a small set of votes, and a classification accuracy of 99.1%. These metrics satisfy the requirements of most civilian and military applications.**

***Keywords*— UAV, audio processing, threat detection, target classification, MFCC, KNN.**

I. Introduction

Constant development of new applications for Unmanned Aerial Vehicles (UAV) attracts the attention of sectors such as law-enforcement [1], [2], agriculture [3], emergency services response [4], and the military [5]. This increasing interest can be attributed to its simplicity and cost-effectiveness. A significant concern, however, is the fact that these devices can also be easily used for illegal activities, such as terrorism [6], unauthorized surveillance [7], and multiple types of cyber-attacks such as eavesdropping, jamming, and spoofing [8]. Therefore, protective measures against such threats, including the capability to detect drones, are necessary to ensure security for critical infrastructures. The task of detecting and classifying drones can be achieved through the processing of signals from a variety of sensors [9], ranging from passive radars [10], [11], optical sensors [12], acoustic sensors [13], to wireless sensors when detecting cyber-space intrusion [14]. This work proposes an acoustic-based method for the detection and classification of drones.

A recent review on auditory perception for unmanned aerial vehicles [13] proposes three categories for detection and classification of a UAV, namely air to land, land to air, and air to air. It also identifies the different purposes of UAV auditory perception in detection and classification of acoustic events, and in source localization. The authors in [13] also refer to the use of microphones as additional tools in the development of new autonomous navigation strategies for UAVs as an open challenge. An example of a non-cooperative collision-avoidance system that uses two microphones to estimate the detection range among aircrafts is described in [15].

Low signal-to-noise-ratio (SNR) can be a problem in detecting and classifying sound events recorded from a drone, mainly because of ego-noise, the noise produced by the drone. This problem can be addressed using classical signal processing noise reduction algorithms, including frequency-spatial filtering techniques, effective in blind source separation problems [16]. More recently, methods based on Deep Neural Networks (DNN) [17] are also being used to enhance speech signals captured using drones. An example is the work presented in [17] that integrates single- and multi-channel DNN-based approaches for the enhancement of speech signals captured from drones.

The acoustic detection can be viewed as an optimization problem that is subject to uncertainty [18] cause by the ego-noise. This uncertainty is similar to that of shooter localization systems that have difficulties differentiating gunshot from firework and other impulsive broadband signals [19]. In the case of the particular problem at hand, it is challenging to discriminate between the ego-noise from the drone carrying the microphone from signals from a potential target drone.

Uncertainty is a matter of concern in any acoustic event classification, especially when the signal-to-noise-ratio is low. It is quite natural that humans need some time sensing the acoustic environment to detect or classify an event among all possible choices. Although subject to this uncertainty, the human auditory system has high accuracy in classification tasks. Thus, we employed the majority voting rule in our system inspired by the human auditory system characteristic; this choice optimizes the capability of classifying an acoustic event as inconclusive when the SNR is low.

The main contribution of our method to the drone clas-

Rigel P. Fernandes and José A. Apolinário Jr. are with the Program of Defense Engineering, Military Institute of Engineering (IME), Rio de Janeiro-RJ, Brazil, E-mails: rigelfernandes@gmail.com, apolin@ime.eb.br. Antonio L. L. Ramos is with the Department of Science and Industry Systems (IRI), University of South-Eastern Norway (USN), Kongsberg, Norway, E-mail: antonio.ramos@usn.no. José M. de Seixas is with the Signal Processing Lab, COPPE/Poli, Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro-RJ, Brazil, E-mail: seixas@lps.ufrj.br. This work was partially supported by the Brazilian Agency CAPES (Project BRA-NORTECH/UTSFORSK, Process no. 23038.018065/2018-17; and Project Joint Passive Coherent Location in 5G and IoT for Critical Infrastructure Protection, Process no. 88881.371305/2019-01), the Norwegian Agency for International Cooperation and Quality Enhancement in Higher Education, Diku (UTF-2018-CAPES-Diku/10002), CNPq, FAPERJ and the Brazilian Navy.

sification problem is the incorporation of the majority voting rule [20] to decide whether to perform a detection and classification round for a potential incoming drone, or to discard the estimate when the voting process turns out to be inconclusive. This simple yet effective approach has a positive impact on minimizing the probability of false positives and false negatives, resulting in a more robust system. Other works available in the literature report high accuracy for the detection task, however lacking a method to deal with the issue of low SNR in parts of the signal [21].

Our method allows the machine learning algorithms to have human-like uncertainty behavior, which prevents estimations when the SNR is too low; that is to say, the model avoids detecting and classifying parts of the signal that are very difficult to perceive as target signals, thereby reducing the probability of false positives and false negatives.

The rest of this paper is organized as follows. Section II describes the techniques used for features extraction, the machine learning algorithm used for acoustic target detection and classification, and the proposed method. Section III provides an overview of the database and a discussion of the experimental results. Conclusions are addressed in Section IV.

II. Target Detection and Classification

A. Problem Statement and Assumptions

In this work, we try to solve the problem of detecting and classifying a single target [22] using audio signals collected from a drone, using a single microphone. The receiver device used to collect the signals in the AIRA-UAS *corpus* [23] is connected of a microphone array aboard a DJI Matrice 100 drone. The targets are additional drones of two different models flying in the near field of the receiver. For the target detection and classification problem, we define three classes as follows. Class 1, which characterizes the absence of drones flying in the near field of the microphone array, Class 2 that indicates a drone 3DR as target and flying near the microphone array, and Class 3 that refers to a Parrot Bebop 2 drone flying in nearby the receiver.

Therefore, the main problem is to detect drones with the presence of the receiver's ego-noise, which shares acoustic features with the target signals.

The signals used in this work, available in [24], are of four different types, namely:

- Background noise $n_1[k]$;
- Ego-noise $n_2[k]$;
- Drone noise emitted by a 3DR Solo drone $s_1[k]$; and
- Drone noise emitted by a Parrot Bebop drone $s_2[k]$.

Class 1 is composed of signals $x_1[k] = n_1[k]$ or $x_1[k] = n_1[k] + n_2[k]$. Class 2 can be formed by signals $x_2[k] = n_1[k] + s_1[k]$ or $x_2[k] = n_1[k] + n_2[k] + s_1[k]$. Finally, Class 3 is composed of signals $x_3[k] = n_1[k] + s_2[k]$ or $x_3[k] = n_1[k] + n_2[k] + s_2[k]$.

B. Feature Extraction

The task of target detection and classification requires features that need to be extracted from acoustic signals that are usually corrupted by noise. From the myriad of feature extraction algorithms [25] available in the literature, we use two that have been employed successfully in related acoustic classification applications, namely the Mel-frequency cepstrum (MFC) [26] and the short-time Fourier transform (STFT) [27].

The basic steps to extract STFT features are as follows. We divide the signals of each class into frames of 960 samples, corresponding to 20 ms at a sampling rate of 48 kHz, with 50% overlap. To reduce spectral leakage, a Hamming window is applied prior to the computation of the 1024-point Discrete Fourier Transform (DFT). We set the window length to 20 ms because we assume the signals to be stationary over that time interval. Most speech signal processing applications make similar assumption. A Fast Fourier Transform (FFT) is applied to the each newly available frame, and the first 513 frequency bins, corresponding to the frequency range $0 \leq \omega \leq \pi$, are used to compose feature vector $\mathbf{X}$ that is used to train the machine learning algorithms. The set of feature vectors for n consecutive frames is represented by data matrix $\mathbf{X}_{\text{STFT}} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\}$.

To extract the Mel-frequency cepstral coefficients (MFCC), we need to retain the logarithm of the amplitude spectrum after smoothing the spectrum and emphasizing perceptually meaningful frequencies. This is achieved using a triangular filter bank to map actual frequency into the Mel scale, a perceptive scaling that better characterizes a sound as perceived by the human auditory system. This mapping is linear below 1 kHz and logarithmically spaced above. The last step, responsible for reducing the number of bins and for transforming the coefficients back to the time domain, is the Discrete Cosine Transform (DCT), which yields 13 coefficients [28]. We also appended the log energy as the first coefficient. The resulting 14×1 vector, $\mathbf{x}_i$, is the MFC feature vector that will also be used in the machine learning algorithms. Data matrix $\mathbf{X}_{\text{MFC}} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ is the set of all MFC feature vectors.

C. The Target Detection and Classification Method

The work in [21] used Convolutional Neural Networks to classify drones based on spectrograms of the signals from the AIRA-UAS database, with a reported result of 97% of detection accuracy. In this work, we used the STFT and the MFC coefficients to build data feature matrices $\mathbf{X}_{\text{STFT}}$ and $\mathbf{X}_{\text{MFC}}$. We then used the corresponding labels, $\mathbf{y}$, to train a KNN classifier and evaluate how the majority voting rule can improve the results. The estimates $\hat{y}_i$ are then used in the majority voting rule strategy, which is adopted to prevent wrong estimations when the signals from three classes are strongly correlated due to ego-noise. The last Δ estimations are the votes. An estimate is only performed if the majority agrees with a certain class. This implementation emulates possible inconclusive estimates that a human being might perform when sensing the acoustic environment. In the validation stage, we used the 10-fold cross-validation strategy.

We tested the majority voting strategy for different Δ frames, $\Delta = \{1, 3, 5, \ldots, 201\}$. The voting interval, t, is calculated as $t = 0.01\Delta + 0.01$ given the 50% overlap between

consecutive frames. When $\Delta = 1$, there is only one vote, which means no voting is performed. When $\Delta = 201$ we have 201 votes, and a decision is made only if 101 votes favor a specific class. The time interval for the first decision is 2.02 seconds, equivalent to 201 frames. Following decisions are made as new frames are available, and are based on the most recent Δ frames. The voting process to make a final decision then requires the current classification plus the 200 most recent classifications. Therefore, the proposed method has a very quick response time and is suitable for real-time use because only one classification is required for each newly available frame.

III. Experimental Results and Discussion

A. Database

We conducted experiments using the AIRA-UAS database [23], [24]. This database is freely available for the scientific community interested in drone noise cancellation, drone detection and classification, and acoustic signal array processing using microphones. It consists of 21 recordings from three different drones; one of them, the receiver drone (DJI Matrice 100), was equipped with eight microphones and we used only one signal, from channel 1. The first set of recordings (Protocol 1) explored different kinematics of the receiver drone carrying the microphones and recorded only background noise or background-noise and ego-noise. Protocols 2 (drone 3DR Solo) and 3 (drone Parrot Bebop 2) recorded the target flying near the array and explored different kinematics. We chose this database due to its availability, different kinematics that produce different spectra, and a good number of recordings. More details are available in [24], [23], [13].

Class 1 has 15 recordings (363 seconds), Class 2 and Class 3 have 3 signals available each (respectively, 85 and 94 seconds in total). We balanced the number of frames of each class for each experiment. Therefore, we used 9 signals in Experiment 1 and 10 signals in Experiment 2.

Figure 1 depicts the spectrograms using the STFT of the signals used in both experiments. Figure 1 (a-f) are related to Class 1, Figure 1 (a-c) are examples of signals containing only background noise, $x_1[k] = n_1[k]$, and Figure 1 (d-f) are examples of signals, $x_1[k] = n_1[k] + n_2[k]$, collected when the receiver drone (DJI Matrice 100) was on the ground with the propellers activated (producing ego-noise), in displacement from the ground to 3 m height, and hovering 5 m height, respectively. Figures 1 (g-i) are related to Class 2. Figure 1 (j-l) are related to Class 3. It should be noted that Figures 1 (f), (i), and (l) hold a strong correlation because the receiver drone was performing the same maneuver (hovering). The only difference is that, in (i) and (l), an additional drone noise (the target signal, $s_1[k]$ or $s_2[k]$, 3DR and Parrot Bebop 2 noises, respectively) is summed to background noise and ego-noise, resulting in $x_i[k] = n_1[k] + n_2[k] + s_i[k]$.

B. Experiments

We performed two experiments to evaluate the proposed method, using different subsets of signals. In Experiment 1, we used signals assumed uncorrelated with the targets; see Figure 1 (a-c). In Experiment 2, we used signals assumed more correlated due to the presence of ego-noise, Figure 1 (d-f). Class 2 is composed of signals depicted in Figure 1 (g-i) and Class 3 is formed by the signals as in Figure 1 (j-l).

Careful examination suggests that the signals depicted in Figure 1 (a-c) are likely to have features with less similarity from the signals of Classes 2 and 3. This level of uncorrelation may yield the best detection and classification results from this *corpus*. The signals of Class 1 in Experiment 2, as illustrated in Figure 1 (c-f), are likely to have features with strong similarity with the signals of Classes 2 and 3, specially the signals depicted in Figure 1 (d-f). This will degrade the results since Class 1 possibly shares many acoustic features with the other classes. Class 1 is composed of 7,563 feature vectors in Experiment 1 and 10,878 in Experiment 2. Class 2 comprises 8,695 feature vectors while Class 3 contains 9,534 feature vectors, in both experiments.

C. Results

We compared the probability of false positive, P_{FP}, and probability of false negative, P_{FN}, for each experiment. The P_{FP} of the cross-validated STFT in Experiment 1 using $\Delta = 1$, according to Figure 2 (a), is close to 0%. The MFCC in Experiment 1 presented the average P_{FP} below 5%.

The results of Experiment 2 using $\Delta = 1$ are depicted in Figure 2 (a), in Experiment 2 STFT also presented better results compared to MFCC. The STFT P_{FP} is equal to 15.2% with k=1 and decays when k increases, except for $k = 2$. In Experiment 2, the MFCC presented the average $P_{FP} = 18.2\%$.

A reliable detection system should have low probability of false positive P_{FP} and false negative P_{FN}, close to 0%. Based on these two experiments, we can conclude that frames of 20 ms are not sufficient to yield satisfactory results when signals are buried in ego-noise. Therefore, we use the majority voting rule as an attempt to minimize P_{FP} and P_{FN} of the experiments.

The STFT-KNN probability of false positive in Experiment 1 was close to 0%, thus we applied our method in Experiment 1 to optimize only the results of MFCC feature vectors. The results can be seen in Figure 2 (b-d). Figure 2 (b) depicts the accuracy (number of accurate estimations of a given class/number of frames of a given class) of the three classes varying Δ with $k = 1$ and Figure 2 (c) with $k = 2$. Table I presents the confusion matrix of MFCC in Experiment 1 with $k = 2$ and $\Delta = 201$. In Figure 2 (d) $k = 20$, we can observe that high values of k creates bias and the performance of our method is reduced, i.e., the accuracy of Classes 1 and 3 is degraded and only Class 2 is enhanced.

TABLE I
Confusion Matrix of Experiment 1 using MFCC with $k = 2$ and $\Delta = 201$.

		Predicted		
	Class	1	2	3
Actual	1	6,473 (100%)	0	0
	2	0	8,026 (100%)	0
	3	0	365	8,495 (95.7%)

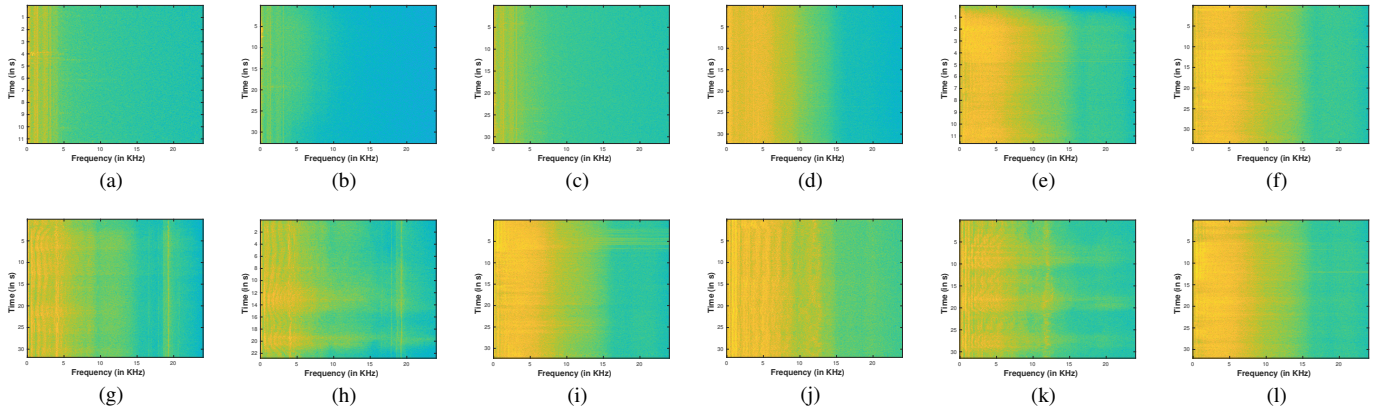

Fig. 1. Spectrograms of the signals used in this work (from channel 1). (a) Class 1 recording 1 (b) Class 1 recording 2 (c) Class 1 recording 3 (d) Class 1 recording 4 (e) Class 1 recording 5 (f) Class 1 recording 6 (g) Class 2 recording 1 (h) Class 2 recording 2 (i) Class 2 recording 3 (j) Class 3 recording 1 (k) Class 3 recording 2 (l) Class 3 recording 3.

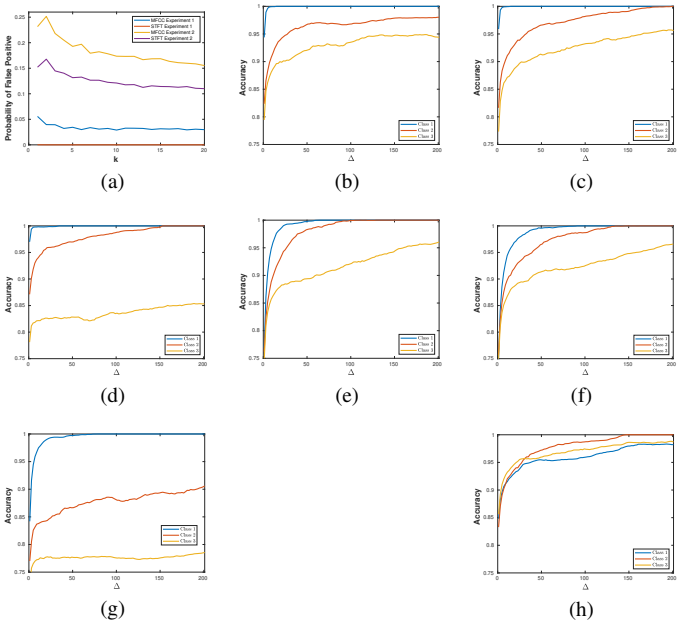

Fig. 2. Classification based on voting schemes using Δ frames. **(a)** KNN with k varying **(b)** MFCC-KNN with $k = 1$ **(c)** MFCC-KNN with $k = 2$ **(d)** MFCC-KNN with $k = 20$ **(e)** MFCC-KNN with $k = 1$ **(f)** MFCC-KNN with $k = 2$ **(g)** MFCC-KNN with $k = 20$ **(h)** best STFT-KNN result.

In Experiment 2, the STFT P_{FP} was lower than MFCC. However, the best results with the voting scheme was achieved with MFCC feature vectors, as we can note in Figure 2 (e-g). The best results in Experiment 2 with MFCC is achieved when $k = 2$; for $k \geq 3$ the performance of Classes 2 and 3 degrades. In Experiment 2, Figure 2 (h) depicts the best results using STFT feature vectors, the P_{FP} of STFT is approximately 3%. Table II presents the confusion matrix best results of Experiment 2 using MFCC ($k = 20$ and $\Delta = 1$).

Table III presents the confusion matrix of MFCC in Experiment 2. Differently from the classical machine learning estimation, the majority voting rule achieved best results when $k = 2$. Class 1 has $10,878 - \Delta + 1 = 10,678$ possible estimates due to the parameter $\Delta = 201$. However, only 8,112 estimates were performed because 2,566 were inconclusive. Likewise, Class 2 has 2,048 inconclusive estimates, whereas Class 3 has 2,546. As can been seen in Table III the individual classification performance of classes 1, 2, and 3 are 100%, 100%, and 97.2%, respectively. Therefore, the classification method using these parameters has an overall accuracy of 99.1%, in average. Discarding the estimates of frames with low SNR leads to performance improvements as illustrated in Table II and Table III.

TABLE II
CONFUSION MATRIX OF EXPERIMENT 2 USING MFCC WITH $k = 20$ AND $\Delta = 1$.

Class	1	2	3
1	9,220 (84.8%)	930	728
2	1,335	6,734 (77.4%)	626
3	1,620	1,000	6,914 (72.5%)

TABLE III
CONFUSION MATRIX OF EXPERIMENT 2 USING MFCC WITH $k = 2$ AND $\Delta = 201$.

Class	1	2	3
1	8,112 (100%)	0	0
2	0	6,411 (100%)	0
3	0	198	6,788 (97.2%)

D. Discussion

The majority-voting rule based on a small set of frames yields better results than algorithms make decisions on frame-by-frame basis. The voting scheme is very straightforward and efficient in reducing false positives and false negatives, thereby resulting in an improved classifier. The MFC feature vector suites well to the acoustic drone detection and classification problem. The results suggest that compressing and mapping the STFT bins onto frequencies better perceived by humans rather than using $\mathbf{X}_{\text{STFT}}$ brings forth MFC feature vectors that are more uniformly distributed in the hyperspace. Moreover,

the compression of STFT feature vectors produces observations with less degrees of freedom, which is beneficial to the majority voting rule scheme. This leads to the conclusion that not all STFT coefficients are related to the noise produced by the drone. The MFC coefficients emulate the confusion that humans make when classifying acoustic events under low SNR. This attribute of the MFC contributes to avoiding bias, and our method is likely to yield better estimations under low SNR.

IV. Conclusions

This paper presented a voting scheme method to optimize the detection and classification of drones using acoustic signatures from a freely available AIRA-UAS drone noise database. The use of STFT as feature vectors yielded the best results initially. The classification accuracy using these features were further improved using our method. We noticed, however, that these features led to biased estimations. Further, we devised an unbiased estimator based on MFCC features combined with the K-Nearest Neighbors algorithm with $k = 1$ or $k = 2$, that performed well even on signals corrupted by strong ego-noise. As future work, we consider using signal enhancement methods to improve the detection besides lifelong algorithms to detect drone classes not used in the training phase.

References

[1] Felipe Gonçalves Serrenho, José Antonio Apolinário Jr., António Luiz Lopes Ramos, and Rigel Procópio Fernandes, "Gunshot airborne surveillance with rotary wing UAV-embedded microphone array," *Sensors*, vol. 19, no. 19, pp. 4271, 2019.

[2] Rigel Procópio Fernandes, Angelo M. C. R. Borzino, António L. L. Ramos, and José Antonio Apolinário Jr., "Investigating the potential of UAV for gunshot DoA estimation and shooter localization," in *Simpósio Brasileiro de Telecomunicações e Processamento de Sinais*. SBrT, 2016, pp. 383–387.

[3] Telmo Adão, Jonáš Hruška, Luís Pádua, José Bessa, Emanuel Peres, Raul Morais, and Joaquim Joao Sousa, "Hyperspectral imaging: A review on UAV-based sensors, data processing and applications for agriculture and forestry," *Remote Sensing*, vol. 9, no. 11, pp. 1110, 2017.

[4] Soumya Kanti Datta, Jean-Luc Dugelay, and Christian Bonnet, "IoT based UAV platform for emergency services," in *International Conference on Information and Communication Technology Convergence (ICTC)*. IEEE, 2018, pp. 144–147.

[5] Marko Suojanen, Vesa Kuikka, Juha-Pekka Nikkarila, and Jari Nurmi, "An example of scenario-based evaluation of military capability areas. An impact assessment of alternative systems on operations," in *Annual IEEE Systems Conference (SysCon) Proceedings*. IEEE, 2015, pp. 601–607.

[6] Gregory D Koblentz, "Emerging technologies and the future of CBRN terrorism," *The Washington Quarterly*, vol. 43, no. 2, pp. 177–196, 2020.

[7] Vinay Chamola, Pavan Kotesh, Aayush Agarwal, Navneet Gupta, Mohsen Guizani, et al., "A comprehensive review of unmanned aerial vehicle attacks and neutralization techniques," *Ad Hoc Networks*, p. 102324, 2020.

[8] Chao Li, Yan Xu, Junjuan Xia, and Junhui Zhao, "Protecting secure communication under UAV smart attack with imperfect channel estimation," *IEEE Access*, vol. 6, pp. 76395–76401, 2018.

[9] Bilal Taha and Abdulhadi Shoufan, "Machine learning-based drone detection and classification: State-of-the-art in research," *IEEE Access*, vol. 7, pp. 138669–138682, 2019.

[10] Yevhen Chervoniak, Rustem Sinitsyn, and Felix Yanovsky, "Passive acoustic radar system for flying vehicle localization," in *23rd International Microwave and Radar Conference (MIKON)*. IEEE, 2020, pp. 278–281.

[11] Marcelo Nogueira de Sousa, Ricardo Sant'Ana, Rigel Procópio Fernandes, Julio Cesar Duarte, José Antonio Apolinário Jr., and Reiner Thomä, "Improving the performance of a radio-frequency localization system in adverse outdoor applications," *Journal on Wireless Communications and Networking*, vol. 1, pp. 123, may 2021.

[12] Yuni Zeng, Qianwen Duan, Xiangru Chen, Dezhong Peng, Yao Mao, and Ke Yang, "UAVData: A dataset for unmanned aerial vehicle detection," *Soft Computing*, pp. 1–9, 2021.

[13] Jose Martinez-Carranza and Caleb Rascon, "A review on auditory perception for unmanned aerial vehicles," *Sensors*, vol. 20, no. 24, pp. 7276, 2020.

[14] Hichem Sedjelmaci, Sidi Mohammed Senouci, and Nirwan Ansari, "A hierarchical detection and response system to enhance security against lethal cyber-attacks in UAV networks," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 48, no. 9, pp. 1594–1606, 2017.

[15] Brendan Harvey and Siu O'Young, "Acoustic detection of a fixed-wing UAV," *Drones*, vol. 2, no. 1, pp. 4, 2018.

[16] Lin Wang and Andrea Cavallaro, "A blind source separation framework for ego-noise reduction on multi-rotor drones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2523–2537, 2020.

[17] Lin Wang and Andrea Cavallaro, "Deep learning assisted time-frequency processing for speech enhancement on drones," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2020.

[18] Nikolaos V Sahinidis, "Optimization under uncertainty: state-of-the-art and opportunities," *Computers & Chemical Engineering*, vol. 28, no. 6-7, pp. 971–983, 2004.

[19] Sami Ur Rahman, Adnan Khan, Sohail Abbas, Fakhre Alam, and Nasir Rashid, "Hybrid system for automatic detection of gunshots in indoor environment," *Multimedia Tools and Applications*, pp. 1–11, 2020.

[20] Razieh Pourdarbani, Sajad Sabzi, Mario Hernández-Hernández, José Luis Hernández-Hernández, Ginés García-Mateos, Davood Kalantari, and José Miguel Molina-Martínez, "Comparison of different classifiers and the majority voting rule for the detection of plum fruits in garden conditions," *Remote sensing*, vol. 11, no. 21, pp. 2546, 2019.

[21] Aldrich A Cabrera-Ponce, J Martinez-Carranza, and Caleb Rascon, "Detection of nearby UAVs using CNN and spectrograms," in *International Micro Air Vehicle Conference and Competition (IMAV)(Madrid)*, 2019.

[22] L Cinelli, G Chaves, and M Lima, "Vessel classification through convolutional neural networks using passive sonar spectrogram images," in *Simpósio Brasileiro de Telecomunicações e Processamento de Sinais*. SBrT, 2018.

[23] Oscar Ruiz-Espitia, Jose Martinez-Carranza, and Caleb Rascon, "AIRA-UAS: An evaluation corpus for audio processing in unmanned aerial system," in *2018 International Conference on Unmanned Aircraft Systems (ICUAS)*. IEEE, 2018, pp. 836–845.

[24] Caleb Rascon, Oscar Ruiz-Espitia, and Jose Martinez-Carranza, "On the use of the AIRA-UAS corpus to evaluate audio processing algorithms in unmanned aerial systems," *Sensors*, vol. 19, no. 18, pp. 3902, 2019.

[25] Thomas F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, Prentice Hall, 2001.

[26] Rigel Procópio Fernandes and José Antonio Apolinário Jr., "Underwater target classification with optimized feature selection based on genetic algorithms," in *Simpósio Brasileiro de Telecomunicações e Processamento de Sinais*. SBrT, 2020.

[27] William Soares Filho, José M. de Seixas, and Natanael Nunes de Moura, "Preprocessing passive sonar signals for neural classification," *IET Radar, Sonar & Navigation*, vol. 5, no. 6, pp. 605–612, 2011.

[28] Beth Logan et al., "Mel frequency cepstral coefficients for music modeling.," in *Ismir*. Citeseer, 2000, vol. 270, pp. 1–11.