UAV Detection and Identification using Acoustic Features and Deep Learning

1st Apurva Anand
Department of Computing Sciences
TAMUCC
Corpus Christi, TX, USA
aanand@islander.tamucc.edu

2nd Ravi Gadgil
Department of Computer Science
San Jose State University
San Jose, CA, USA
ravi.gadgil@sjsu.edu

3rd Helen Lei
College of Computing and Information Science
Cornell University
Ithaca, New York
hl883@cornell.edu

4th Sandeep Amgothu
Department of Computing Sciences
TAMUCC
Corpus Christi, TX, USA
samgothu@islander.tamucc.edu

5th Dulal Kar
Department of Computing Sciences
TAMUCC
Corpus Christi, TX, USA
Dulal.Kar@tamucc.edu

***Abstract*—Unmanned aerial vehicles (UAVs), or drones, offer immense potential but pose increasing security concerns due to their accessibility and misuse. Detection and identification of drones is extremely important in the mitigation of these risks. This work explores the application of different preprocessing techniques and deep learning models to the identification of drones by their acoustic signature. Key findings include the effectiveness of deep learning models in achieving promising results for audio identification of drones. However, the success of deep learning relies heavily on a large dataset. To address this, this work will also explore combining various datasets and data augmentation. In conclusion, this analysis compares the immense, varying potential of utilizing different acoustic features and deep learning algorithms for accurate, real-time UAV identification.**

***Index Terms*—drone, UAV, drone audio, drone identification, deep learning, CNN, RNN, CRNN**

I. Introduction

In recent years, drones have transcended their traditional military applications and have become pervasive across various industries and recreational pursuits. From facilitating last-mile goods delivery to revolutionizing agriculture, surveying, emergency response, and cinematography, drones have rapidly integrated into commercial and commodity sectors [6]. The global drone market is poised for exponential growth, forecasted to surge to 13 billion US dollars by 2025, marking a remarkable 200% expansion within a mere five years [7]. Although some nations have imposed bans on drone sales and usage, many countries facilitate straightforward drone acquisition, subject to compliance with stipulated regulations [8]. However, the widespread availability of drones also brings forth concerns regarding potential misuse, ranging from illicit activities such as drug smuggling to spying on sensitive locations such as military sites or civilian homes. In light of escalating reports documenting such incidents in recent years, there is an urgent need to explore effective methods for drone identification to ascertain if the UAV in a particular airspace is authorized or hostile. Detecting and identifying flying drones poses a significant challenge due to their high speed and diverse shapes, especially when they are small in size and traditional detection methods, such as RF or optical sensors, are used. However, different drone models often exhibit distinctive acoustic signatures, presenting an opportunity to enhance identification systems through UAV acoustic recognition technology. Audio signals contain different types of features such as Mel-frequency cepstral coefficients (MFCCs) and short time Fourier transforms (STFTs). The values of these signal representations vary for the sounds that are generated by different types of UAVs and other audio sources. Consequently, these features can be leveraged to train Deep Learning (DL) neural networks in order to distinguish between drone and non-drone signals, and even identify specific drone models. Then, the performance of models using MFCCs versus STFTs can be compared to establish which feature type is more effective for drone identification. This approach not only addresses current challenges in drone management, but also lays groundwork for advancing drone detection capabilities amid evolving technological landscapes.

National Science Foundation (NSF)

II. Related Works

Several methods of drone detection have been studied in prior research, including radar, LiDAR, and computer vision. However, these methods suffer from various shortfalls that hinder their efficacy in real-world scenarios. For example, radar excels when detecting large objects such as aircraft and missiles [9]. However, this isn't as useful when attempting to identify smaller and more maneuverable objects like commercial drones that can be easily adapted for nefarious uses. In addition, LiDAR is highly dependent on the weather conditions, so it won't be as effective when faced with turbulent environmental factors [10]. Similarly, computer vision methods are significantly hampered by adverse weather due to limited visibility. Also, they have difficulty in congested urban

environments where numerous objects in a small location obscure drone detection [11]. These factors underscore the need for an alternative detection technology that is capable of overcoming these specific challenges in diverse operational environments.

Audio detection is a newer method that holds a lot of potential and can be relatively cheaper then methods like radar [1]. This proposal will focus on drone detection through audio processing. There are many different features that can be extracted from audio data. The process involves using a Hamming window to reduce edge effects, then extracting various features from each subframe of the windowed signal. These features include Short-Time Energy, Temporal Centroid, Spectral Centroid, Spectralroll-off, and Mel frequency coefficients. By analyzing their acoustic signatures, they are incredibly useful in UAV detection. [2]

With the goal of optimizing the detection of drones through audio, many previous works have explored the extraction of Mel-frequency cepstral coefficients (MFCC) features as well as short time Fourier transform (STFT) in processing drone audio clips. Study [3] compares the effectiveness of each of these methods in combination with SVM and CNN classification models. They also tested different configurations of listening nodes. They found that MFCC-SVM was unable to distinguish between plane and drone sounds well and the best configuration uses STFT-SVM. CNN didn't work that well for this task since only binary classification and a lot of resources were needed to train CNN.

Another work [4] compares a STFT-based CNN approach to GMM and RNN models using MFCCs and mel-spectrograms on a dataset that included various background noises, and found the STFT-based CNN achieved lower false positive rates for signals similar to drones, further highlighting the effectiveness of the use of STFT features and a CNN model for audio drone detection in complex environments.

With the use of deep learning models, a large dataset is required. Addressing the scarcity of drone sound data, the study in [5] explores data augmentation to synthesize realistic training data. This is helpful as many works, such as [3], faced limitations as they struggled with data collection.

III. Proposed Methodology

This project consists of four main phases: data collation, preprocessing and feature extraction, implementation and training of deep learning models using the dataset, and evaluation of models using baseline comparisons and various performance metrics.

A. Data Acquistion

The scarcity of drone audio data poses a significant challenge in UAV acoustic detection tasks employing Deep Learning. Consequently, we have to make extensive efforts to synthesize a dataset of an adequate size to train the models. Here are the four main avenues that we will be using to collate data:

Experimentally-Captured Drone Sounds: Utilizing the built-in microphone feature of the iPhone 13, we will meticulously record drone acoustic profiles, covering a spectrum of drone classes including CrazyFlie, DJI Phantom 4 Pro, and DJI Tello. Flying drones are carefully orchestrated at 1 meter above ground within a 1 meter radius.

YouTube-Sourced Drone Sounds: The dataset will be augmented further with drone acoustic profiles sourced from YouTube videos. This segment encompasses a diverse array of drone models, including DJI MavicPro, DJI Matrice 100, and DJI Spark.

Online Databases-Sourced Drone Sounds: Supplementary drone acoustic profiles are curated from reputable online databases, featuring models like Parrot Bebop and Parrot Membo.

Non-Drone Sounds: To ensure dataset balance, non-drone acoustic profiles will be collected both experimentally and from online databases. These audio clips will capture ambient sounds such as airplane noise, bird chirps, calm environments, human speech, rainfall, and traffic.

This meticulous dataset compilation process ensures the model's robustness and ability to generalize across various drone models and acoustic environments.

B. Data Preprocessing

- **Short-time Fourier transform (STFT)** is one method of feature extraction in processing audio data. The audio is segmented and then each segment is processed separately to compute the STFT and extract the magnitued. The formula used is given as follows:

$$X(m,k) = \sum_{n=0}^{N-1} x(n+mH) \cdot w(n) \cdot e^{-i\frac{2\pi}{N}nk}$$

  where:
  - $X(m,k)$ is the STFT of the signal.
  - $x(n+mH)$ is the input signal at frame m and sample n, with a hop size H.
  - $w(n)$ is the window function.
  - $e^{-i\frac{2\pi}{N}nk}$ transforms the signal to the frequency domain.
  - N is the FFT window size.

  We plan to initially use a window size of 2048 and a hop size of 512, however we will explore other window and hop sizes for the best performance due to tradeoffs between time resolution and frequency resolution that come with different window sizes. A larger frame size would result in higher frequency resolution but lower time resolution while a smaller frame size would result in lower frequency resolution but lower time resolution. The window function used is the Hann function.
- **Mel-frequency Cepstral Coefficients (MFCCs)** constitute another valuable feature set for building a robust deep learning model. Meticulous data preprocessing techniques are employed to extract this primary feature set which

will later be used to train the models. The formula to compute these coefficients is provided here:

$$MFCC_n = \sum_{m=1}^{M} \log |Y_m| \cos \left[2\pi (n-1) \left(\frac{f_m}{f_s} + \frac{k}{2M} \right) \right]$$

where:

- n is the MFCC coefficient number (usually ranges from 1 to N).
- M is the number of mel-frequency bins.
- Y_m is the magnitude spectrum of the signal at the m^{th} mel-frequency bin.
- f_m is the center frequency of the m^{th} mel-frequency bin.
- f_s is the sampling frequency of the audio signal.
- k is a constant value used for warping the frequency scale (often set to 0).

MFCC Extraction: MFCCs are renowned for their effectiveness in capturing essential characteristics of audio signals. They are extracted from the raw audio files using the librosa library. This process will facilitate MFCC computation with optimal efficiency and accuracy. Key parameters governing the MFCC extraction process include:

- **Sample Rate:** The audio sample rate, set at 22050 Hz, ensures adequate representation of the audio signals.
- **Duration:** Audio clips, each spanning 10 seconds, provides ample temporal context for feature extraction.
- **Number of MFCC Coefficients:** A total of 20 MFCC coefficients will be computed to encapsulate relevant spectral information.
- **FFT Window Size and Hop Length:** The FFT window size, set at 2048, and the hop length of 512 frames are pivotal in partitioning the audio signals into smaller segments for MFCC computation.
- **Number of Segments per Audio Clip:** Each audio clip is segmented into 10 segments to facilitate fine-grained analysis and feature extraction.

Labeling: Appropriate labeling of the dataset is paramount to enable supervised learning and model evaluation. Semantic labels corresponding to distinct drone classes are inferred from the directory structure of the dataset. Each directory name serves as a unique semantic label, ensuring clear delineation of drone classes during model training and testing. Additionally, all of the non-drone audio samples will be labeled with same semantic label.

Data Organization: The dataset is meticulously organized to facilitate seamless processing and subsequent model training. Data structures are defined to accommodate the extracted MFCCs, along with corresponding labels and semantic mappings. Each MFCC vector, derived from individual audio segments, will be serialized to a JSON file, preserving the integrity of the dataset structure. In summary, the data preprocessing pipeline encompasses the extraction of MFCCs as the primary feature set, complemented by meticulous labeling of the dataset. These preparatory steps establish a solid foundation for subsequent model development and validation, facilitating the construction of robust and accurate deep learning models for drone identification.

C. Model Architecture

We will evaluate three different models: Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and Convolutional Recurrent Neural Network (CRNN).

CNN: The specific CNN architecture is a sequential model comprised of the following layers:

- **Convolutional layers**: These layers perform convolutions, which are operations that identify and extract patterns within the audio data. The model utilizes two convolutional layers, each with 8 and 32 filters respectively, and kernel sizes of 5×5. These filters operate with a ReLU (Rectified Linear Unit) activation function.
- **Max pooling layers**: Following each convolutional layer is a max pooling layer with a pool size of 2×2. This layer downsamples the data, reducing its dimensionality while retaining the most significant features.
- **Batch normalization layers**: These layers are inserted after the convolutional layers to address internal covariate shift.
- **Flatten layer**: This layer transforms the extracted features from a 2D format into a 1D vector suitable for feeding into the fully-connected layers.
- **Dense layers**: The model utilizes two fully-connected layers. The first layer has 32 neurons with a ReLU activation function, followed by a dropout layer with a rate of 0.3 to prevent overfitting. The final dense layer has 2 neurons with a softmax activation function, corresponding to the two output class labels.

RNN:

- **Bidirectional recurrent layers**: These layers capture sequential dependencies in the audio files by processing them in the forward and backward directions. This feature helps enhance contextual understanding. The model utilizes two bidirectional recurrent layers which both have 32 units. These filters operate with a ReLU (Rectified Linear Unit) activation function.
- **Batch normalization layers**: These layers are inserted after the recurrent layers to address internal covariate shift.
- **Dense layers**: The model utilizes two fully-connected layers. The first layer has 32 neurons with a ReLU activation function, followed by a dropout layer with a rate of 0.3 to prevent overfitting. The final dense layer has 2 neurons with a softmax activation function, corresponding to the two output class labels.

CRNN: This model combines the CNN and RNN models to synthesize their advantages for enhanced performance. The architecture removes the Dense layers from the CNN. It also adds a Reshape layer in between the two models to modify the dimensions of the input data before passing it into the RNN model.

D. Model Evaluation

The performance of the models will be evaluated using the following metrics: accuracy, F1-score, precision, and recall. Furthermore, we will benchmark our results against established models in the literature, such as those referenced in [12], to gauge performance quality.

IV. Plan

The following itemized list details our goals for each week of the project:

A. Week 5: Replication of Previous Experiments and Midterm Presentation

To address challenges such low accuracy and erratic model behavior for both feature sets, attempt to implement the binary and multi-class classification experiments in [12] using the same dataset, extracted feature set, and models. Develop a midterm presentation that outlines research findings, current challenges, and future work.

B. Week 6: Data Collection and Preprocessing

Compile drone audio clips covering a wide range of UAV models through experimental recordings, YouTube videos, and online databases. Collate non-drone noise clips sourced from online sources. Synthesize these clips into a binary classification dataset by labeling each file as "drone" or "non-drone". Extract the STFT and MFCC feature sets from the audio clips. Serialize the feature sets as feature vectors in JSON files. If the experiments performed in Week 5 have shown decent results, use the preprocessing methods from that implementation.

C. Week 7: Binary Audio Classification - Implementation

Create and train a CNN, RNN, and CRNN model for each individual feature set. If the experiments conducted in Week 5 have shown promising results, begin by using those models as a basis and refine them according to the new dataset. Observe the results and attempt to optimize performance using methods such as normalization, pitch-shifting, or increasing model complexity.

D. Week 8: Multi-Class Audio Classification - Data Preparation and Implementation

Label each JSON file from the binary classification dataset with its appropriate drone model name. Organize these files into folders named according to their semantic labels. Combine these directories to synthesize a new Multi-Class Classification (MCC) dataset. Include the pre-labeled "non-drone" files from the prior experiment in a separate directory named "non-drone" within the new dataset. Create and train a CNN, RNN, and CRNN model for the STFT and MFCC features sets.

E. Week 9: Multi-Class Audio Classification - Model Optimization

Observe the results of the MCC models relative to benchmark models in the literature and attempt to optimize performance using methods such as normalization, pitch-shifting, or increasing model complexity.

F. Week 10: Evaluation, Final Report, and Final Presentation

Evaluate the models using the prescribed performance metrics and compare them to benchmark models from prior research. Compare the performance of the models for drone identification when using STFTs versus MFCCs. Describe the methodology, results, and conclusions of the project in the final report and final presentation.

V. Expected Results

The goal of this research is to analyze the effectiveness of different preprocessing techniques in the use of deep learning models for drone identification. We expect to obtain the most optimal accuracy in multi-class drone classification by comparing the use of STFT features or MFCC coefficients in CNN and RNN models. We also aim to achieve widely generalizable results to improve upon previous models that only work well under certain conditions.

References

[1] M. A. Khan, H. Menouar, A. Eldeeb, A. Abu-Dayya and F. D. Salim, "On the Detection of Unauthorized Drones—Techniques and Future Perspectives: A Review," in IEEE Sensors Journal, vol. 22, no. 12, pp. 11439-11455, 15 June 15, 2022.

[2] A. Bernardini, F. Mangiatordi, E. Pallotti, and L. Capodiferro, "Drone detection by acoustic signature identification," in *Proceedings of the International Conference on UAV Detection and Identification*, Rome, Italy, pp. 45-50, 2022.

[3] B. Yang, E. T. Manson, A. Smith, E. Diatz, and J. Gallagher, "UAV detection system with multiple acoustic nodes using machine learning models," International Journal of UAV Research, vol. 15, no. 3, pp. 123-135, 2022.

[4] Seo, Yoojeong, Beomhui Jang, and Sungbin Im. "Drone detection using convolutional neural networks with acoustic STFT features." 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2018.

[5] Jeon, Sungho, et al. "Empirical study of drone sound detection in real-life environment with deep neural networks." 2017 25th European Signal Processing Conference (EUSIPCO). IEEE, 2017.

[6] Altawy, Riham, and Amr M. Youssef. "Security, privacy, and safety aspects of civilian drones: A survey." ACM Transactions on Cyber-Physical Systems 1.2 (2016): 1-25.

[7] I. Wagner. [n.d.]. Commercial UAVs - Statistics & Facts. arXiv:https://www.statista.com/topics/3601/commercial-uavs/

[8] Jones, Therese. International commercial drone regulation and drone delivery services. No. RR-1718/3-RC. Santa Monica, CA, USA: RAND, 2017.

[9] M. Benyamin and G. H. Goldman, "Acoustic detection and tracking of a class i UAS with a small tetrahedral microphone array."

[10] A. R. Wagoner, D. K. Schrader, and E. T. Matson, "Towards a vision-based targeting system for counter unmanned aerial systems (CUAS)," in 2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), pp. 237–242.

[11] B. Yang, E. T. Matson, A. H. Smith, J. E. Dietz and J. C. Gallagher, "UAV Detection System with Multiple Acoustic Nodes Using Machine Learning Models," 2019 Third IEEE International Conference on Robotic Computing (IRC), Naples, Italy, 2019, pp. 493-498, doi:10.1109/IRC.2019.00103.

[12] Al-Emadi, S.; Al-Ali, A.; Al-Ali, A. Audio-Based Drone Detection and Identification Using Deep Learning Techniques with Dataset Enhancement through Generative Adversarial Networks. Sensors 2021, 21, 4953. https://doi.org/10.3390/s21154953