Robust Drone Detection for Acoustic Monitoring Applications

Mattes Ohlenbusch*, Aike Ahrens*†, Christian Rollwage*, Jörg Bitzer*†
*Fraunhofer Institute for Digital Media Technology; Project group Hearing, Speech and Audio Technology
†Jade Hochschule Wilhelmshaven, Oldenburg, Elsfleth; Department for Technology and Health
Oldenburg, Germany
mattes.ohlenbusch@idmt.fraunhofer.de

***Abstract*—Commercially available light-weight unmanned aerial vehicles (UAVs) present a challenge for public safety, e.g. espionage, transporting dangerous goods or devices. Therefore, countermeasures are necessary. Usually, detection of UAVs is a first step. Along many other modalities, acoustic detection seems promising. Recent publications show interesting results by using machine and deep learning methods. The acoustic detection of UAVs appears to be particularly difficult in adverse situations, such as in heavy wind noise or in the presence of construction noise. In this contribution, the typical feature set is extended to increase separation of background noise and the UAV signature noise. The decision algorithm utilized is support vector machine (SVM) classification. The classification is based on an extended training dataset labeled to support binary classification. The proposed method is evaluated in comparison to previously published algorithms, on the basis of a dataset recorded from different acoustic environments, including unknown UAV types. The results show an improvement over existing methods, especially in terms of false-positive detection rate. For a first step into real-time embedded systems a recursive feature elimination method is applied to reduce the model dimensionality. The results indicate only a slight decreases in detection performance.**

***Index Terms*—Drone detection, UAV, public safety, binary classification, acoustic event detection, feature selection.**

I. Introduction

Unmanned aerial vehicles (UAVs) have become increasingly popular in commercial and private use contexts [1]. Applications are manyfold, such as in (semi-) automated monitoring or photography. As amateur UAVs (often called drones) are now commercially available with little regulations, their uncontrolled deployment may pose a considerable security risk to the public. Past events such as smuggling drugs into prisons, intrusion of government institutions and disturbances of airport traffic show that drones already represent a real threat [2, 3]. The development of robust and reliable detection methods for UAVs can therefore be considered essential for public safety. Although systems for drone detection exist, they come with individual limitations due to their mode of operation as well as restrictions of deployment for a given geographical environment. Depending on the environment, traffic or birds can influence different types of detector heuristics. Often line-of-sight is necessary for the detection.

This contribution focuses on acoustic drone detection methods. Acoustic sensors are very cost effective, and are also less impaired by adverse weather conditions, in contrast to other potential modalities such as video or thermal-based detection [4]. Video-based detection may have problems in darkness, and thermal-based detection is required to account for temperature changes. In addition, UAVs emit very distinct acoustic signatures due to their fast rotating propellers, single- or multi-rotor UAVs have harmonic acoustic emissions, which can be recognized very well by humans [5]. The distinction by humans through hearing alone indicates that by means of acoustic sensing, it is reasonable to detect UAVs as well.

However, despite their assumed suitability for drone detection, acoustic sensors are currently not used in e. g., airport surveillance [2].

II. Related Work

Usually, acoustic drone detection is formulated as a (often binary) classification task. In [6, 7], a correlation-based approach was described. The accuracy of this approach was found to be quite low in a real-time environment [4]. In an effort to explore deep learning architectures for drone classification, Jeon et al. [8] designed and compared Gaussian mixture models (GMM), convolutional neural networks (CNN) and recurrent neural networks (RNN) with the inputs being either mel-spectrogram or mel-frequency cepstral components (MFCCs), which are known to represent spectral characteristics such as periodicity and coloration, and are often applied to acoustic event detection or classification problems [9]. Promising results were reported in particular from the RNN architecture, but also the lack of training data to fully utilize the potential of the data-driven approaches if combined with higher dimensional feature sets. In another recent paper, the authors also considered similar architectures, but found a CNN structure to be more beneficial for drone detection [3]. The slightly artificial dataset considered was made publicly available, and contains recordings of two drones flying indoors as well as white noise, silence and various environmental sounds originating from the online audio database freesound.org, such as animal noises, keyboard tipping, and crackling fire. An approach also using cepstral features (cepstrum values instead of coefficients) in combination with logistic regression was published in [10], but surpassed by the authors in [11], where a CNN-based approach on spectrogram features was utilized. Both systems are based on 20 ms-frames. In contrast, [12] proposed a support vector machine (SVM) classification method

in which short-time features such as MFCCs and various spectral and temporal measures are first extracted in frames of similar lengths. From these, statistical measures like mean and variance are computed. The reason behind this approach is to increase the robustness of the method against acoustic events that share similar spectral characteristics with UAV emissions, but vary in temporal structure. Such events arguably cannot be distinguished in short time frames but only over longer periods. A similar approach also employs MFCC feature vectors in an SVM classification framework [4]. The superiority of this approach over methods, such as [7], is demonstrated on a small dataset. The authors also mention a lack of data for large scale training of deep learning approaches.

In this work, we will compare some of these approaches, extend one of them, and evaluate the systems based on real-world data.

III. Proposed Detection Framework

In order to establish features that represent both spectral and temporal characteristics, Bernardini et al. [12] combine block-wise features from 20 ms-frames (with an overlap of 10 ms) of audio and statistical analysis over several blocks. Block-wise time domain features utilized in this approach are the Short Time Energy (STE), the Temporal Centroid (TC) and the Zero Crossing Rate (ZCR). In the spectral domain, the Spectral Centroid, Spectral Roll-Off and 13 MFCCs are extracted.

Over a period of 200 ms are the mean and the sample variance of the local features are computed over 20 blocks each. The detection is carried out using one-class SVM classification, based on the normalized mid-term statistics for each block-wise feature, instead of using the features themselves. The kernel function used are radial basis functions (RBF), also known as Gaussian kernels.

A. Additional Features

Following this approach, this paper improves the performance of drone detection by the use of additional block-wise features in addition to the feature set used in [12]:

Block RMS: The root-mean-squared value per audio frame, defined as

$$\mathrm{RMS} = \sqrt{\sum_{k=0}^{N-1} s^2(k)}, \tag{1}$$

with frame length of N and a sampled audio signal denoted by s (used instead of STE).

Spectral Bandwidth: Using the Spectral Centroid (SC) definition from [12], the Spectral Bandwidth is computed as

$$\mathrm{SB} = \sqrt{\sum_n p(n) \cdot (f(n) - \mathrm{SC})^2} \tag{2}$$

for an individual audio frame, with frequency index n (up to the Nyquist frequency), the Short Time Fourier Transform (STFT) power p and the STFT frequency f.

Spectral Contrast: In order to represent the relative spectral distribution, the Spectral Contrast [13] in 7 subbands was utilized.

Spectral Flatness: The Spectral Flatness measures how tonal or flat an audio frame's STFT spectrum is [14].

Spectral Flux: Onsets in audio wave forms can be detected from comparisons between two successive STFT magnitude blocks [15]:

$$\mathrm{SF}(b) = \sum_n R\left(|p(b,n)| - |p(b,n-1)|\right), \tag{3}$$

where $R(x) = \frac{x+|x|}{2}$ is the half-wave rectifier function.

Skewness and Kurtosis: Aside from the feature statistics, the time series skewness and kurtosis were included as frame-wise features as well.

Furthermore, the skewness

$$\kappa_j = \frac{m_3}{\sqrt{m_2^3}}, \tag{4}$$

is added as an additional statistical measure where j is the mid-term index and m_i is the ith moment defined as

$$m_i = \frac{1}{N} \sum_{k=0}^{N-1} (\phi_{j+k} - \mu_j)^i \tag{5}$$

using feature values ϕ and their mean μ_j. The reason for using the skewness in addition to mean and variance is to detect skewed feature distributions, which may be a result of very short, impulse-like acoustic events.

In the extended method, the python library *librosa* [16] was used for calculating the majority of the block-wise features.

B. Fitting Procedure

The input data for the proposed method consists of feature vectors of dimensionality $3 \cdot 31 = 93$, since three moments for 31 features dimensions were computed. These vectors are centered and scaled in each dimension, as a pre-processing normalization step. Each normalized feature value is computed as

$$x_{\text{normalized}} = \frac{(x - \mu_{\text{train}})}{\sigma_{\text{train}}}, \tag{6}$$

where x is the original feature value, μ_{train} the mean of this feature in the training dataset, and σ_{train} the respective sample standard deviation.

We split the training data into 70% for the actual fitting, and the remaining data for the validation process. Only a subset of the data is used in individual fits, since each fit is done in ten-fold cross-validation. The training is carried out over a grid search of tuning parameters, with the values for cost $C = \{0.25, 0.5, 1, 2, 4\}$ and kernel scale $\gamma = \{0.001, 0.01, 0.1, 1, 10\}$. The degree of the kernel polynomial was also determined by cross-validation, resulting in a linear kernel.

C. Description of Dataset

The dataset was recorded in two sessions at different places. The training data and the test data were taken from different individual recording environments. The training material contains five different types of drones (DJI Phantom4, DJI Mavic, custom-build racing drone, SKY-HERO Little Spyder, ALIGN M690L Multi-Drone). The testing material contains different types of UAVs that were not present in training dataset recordings (Unique Taifun H520, DJI F450) as well as recordings of a surface plane (Parrot Disco). Additionally, the ALIGN and SKY-HERO also featured in training data were recorded in different outdoor environments for the testing dataset. In total, three UAV types not present in the training data were used for testing. The total recording time of UAV emissions amount to 1.9 h, with 68 % being accounted for by the training data, and the rest by the test dataset.

Additionally, the same amount of non-UAV recordings such as environmental sounds were included in the dataset. Recordings include traffic sounds, construction noise, bird calls, wind and engine sounds. While UAV recordings tend to be very similar in terms of the inherit harmonic structure, environmental noise tends to be more diverse, meaning that not only stationary noise but also harmonic or transient sounds may occur, which we included in the dataset. Examples of harmonic sounds include bird calls or church bells, and construction noise features a varying structure originating from different kinds of tools and machinery, such as drill hammers. Both training and testing data feature a wide range of signal-to-noise ratios due to part of the recordings originating from dynamic acoustic environments, ranging from farmland to urban areas. No audio files were mixed from signal and noise, but rather represent the acoustic situations they were recorded in. Spectrogram representations of some recordings are shown in Fig. 1.

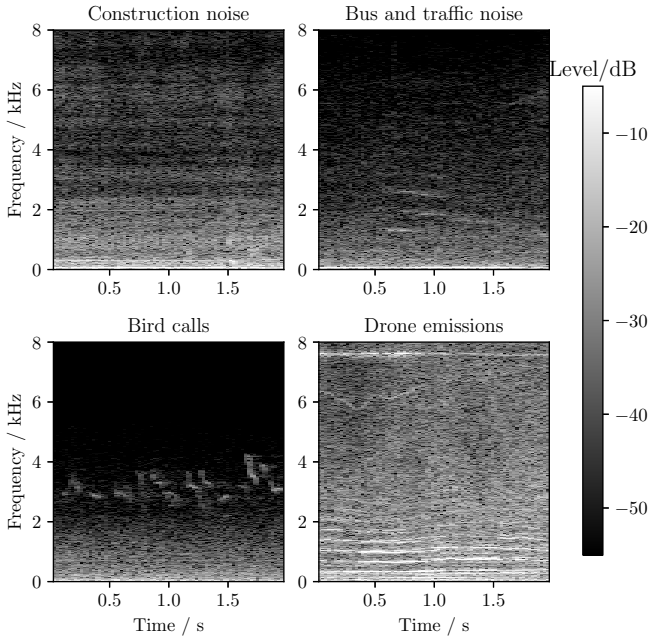

Fig. 1. Spectrograms for different environmental noise, and an UAV.

IV. Experimental Results

In order to evaluate acoustic drone detection methods, it is necessary to estimate their detection performance on unseen audio recordings.

A. Detection Experiment Using Unseen Recordings

The proposed method described in section III was used to detect drones in the unseen dataset, as well as the MFCC-based SVM approach described in [4] and the Spectrogram-based Convolutional Neural Network proposed in [11]. The approach described in [4] was implemented within the same training and cross-validation framework as the proposed method. The CNN-based method from [11] was implemented as well, using an FFT size of $N = 64$ and frame length of 20 ms for the spectrograms. The network was optimized using the stochastic gradient descent algorithm, with a learning rate of 0.0001. Since the dataset used in the original publication was smaller than the one considered, the training was carried out over 5 epochs. Kernels with size 5×5 were utilized, and individual spectrograms were zero-padded before convolution in order to keep the given dimensionality as described for the original architecture. The implementation considered in this experiment was implemented using the PyTorch 1.1.0 Framework [17].

The performance is evaluated in terms of their accuracy, sensitivity and specificity, and their F1-Score. In addition, the false positive rate, which is of high interest in this application, is computed. Acoustic monitoring applications can require especially low false positive rates, because regular false alarms will encourage security staff to ignore UAV detection output.

A confusion matrix for the unseen testing dataset, as classified by the proposed method, is shown in Table I. It is observed that the rate of false positive detection is very low, leading to robust detection results. Additionally, the majority of UAV occurrences is correctly detected. For comparison, condensed results for the unseen testing dataset using different methods are listed in Table II. It can be seen that the SVM-based approaches perform better on the testing dataset than the spectrogram-CNN, which may be due to the fact that the testing dataset contains time-varying and transient noises with momentary harmonic structure, such as church bells or engine sounds. The proposed detection approach takes into account multiple successive time frames, allowing to detect changes in features via mid-term statistics.

TABLE I
Confusion Matrix for the Unseen Testing Dataset, as Classified by the Proposed Method.

		Predicted	
		no UAV	UAV
Actual	no UAV	18021	1580
	UAV	1095	14930

TABLE II
RESULTS OF THE DETECTION EXPERIMENT ON AN UNSEEN DATASET, IN PERCENT.

Method	Acc.	F1-Score	Sens.	Spec.	False Pos.
Spec-CNN [11]	85.68	84.76	79.78	91.56	08.44
MFCC-SVM [4]	88.36	88.85	90.36	86.63	13.37
Proposed (SVM)	92.63	93.19	90.95	94.07	05.92

In order to further investigate this assumption, the class-specific accuracy for some subdivisions of the two classes present in the testing dataset from the experiment were computed. They are listed in Table III. The divisions were chosen after initial testing to illustrate the varying robustness against certain kinds of disturbances and problematic acoustic events in this two-class problem.

TABLE III
RESULTS OF THE DETECTION EXPERIMENT FOR THE PROPOSED METHOD, DIVIDED INTO SUBCLASSES.

Subclass	Subclass accuracy	Frame count
UAV (close/moderate distance)	93.53 %	10820
UAV (quiet/very distant)	78.75 %	4160
UAV + construction noise	91.96 %	1530
Environmental noise	98.35 %	17530
Heavy wind noise	83.66 %	606
Construction noise	32.45 %	530
Church bells	98.00 %	50
Bird sounds	24.50 %	400

It can be seen that for the majority of the test dataset, UAVs can be detected from a lot of noise subclasses. Recordings of more quiet or more distant UAVs were less likely to be detected. From the noise subclasses, construction noise and bird calls were confused by the proposed method very often, resulting in false positive UAV detection. As a possible solution to overcome this limitation, more recordings of those could be integrated into a future training dataset. For both subclasses, the events can potentially occur over multiple successive time frames and exhibit harmonic structure in the same frequency range as UAV emissions.

B. Downsizing the Model by Recursive Feature Elimination

In order for a detection model to be applicable to real-world scenarios with limited resources, it may be required to decrease the dimensionality of the feature space in the classification problem. The decreased computational complexity allows for cost-sensitive solutions and this allows to deploy multiple sensor nodes within the detection perimeter. Additionally, coverage of an event by multiple detectors may further increase the efficiency of detection methods. For these reasons, we reduced the number of features in the model in order to evaluate a low-dimensional variant of the proposed method. For this experiment, recursive feature elimination originally described in [18] was utilized as implemented in the caret R library [19]. The same parameters included in training the method as described in section III-B were used. Training accuracy for the same test procedure as before is shown in Figure 2 for each selected reduced subset. The five-

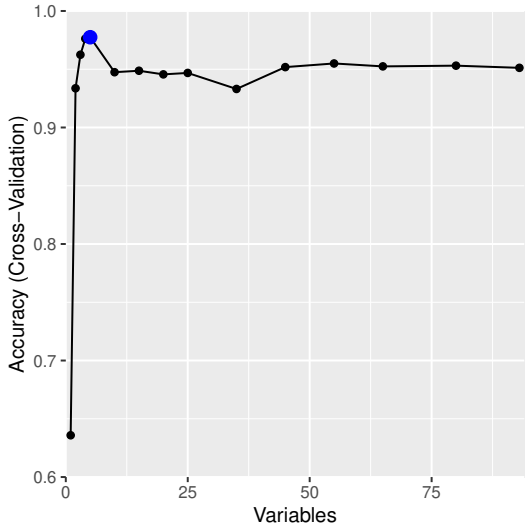

Fig. 2. Accuracy from cross-validated recursive feature elimination on the training dataset, for subsets of the features used in the extended model.

dimensional model (as marked in the plot) was selected and used to detect drone occurrences on the test dataset. Table IV shows the results for the reduced model with the highest accuracy in RFE. The model makes use of the first order statistics of the fourth and eighth MFCC, as well as the spectral contrast in the second, sixth and seventh band. These features are particularly important for UAV fundamental and harmonic overtone power, and detecting abrupt changes in these bands. The results illustrate that comparatively good performance

TABLE IV
RESULTS OF THE DETECTION EXPERIMENT ON AN UNSEEN DATASET IN PERCENT, USING A REDUCED FEATURE SPACE OF $F = 5$ FEATURES.

Method	Acc.	F1-Score	Sens.	Spec.	False Pos.
Redux-SVM	89.26	89.98	88.54	89.88	10.11

is achievable on the testing dataset, but the reduced model dimensionality comes with a considerable increase in false positives. In any practical application, this trade-off between feature dimensionality and false detection rate would need to be addressed.

V. CONCLUSIONS

The experiment with datasets originating from different recording environments resulted in very promising detection performance results for the proposed method. While the detection of close or moderately distant UAVs is very accurate, for more distant or quiet UAVs the performance is slightly impaired. For different kinds of acoustic noise or other non-UAV sounds, the false-positive detection rate was found to

be too high for construction noise, as well as for bird call recordings. It is assumed that incorporating more examples of these noise types into the training data may further increase the classification performance. Since the length of individual occurrences depends on the acoustic environment and UAV velocity, higher mid-term frame lengths may result in a decrease in realistic environments where possible detection time is shorter due to SNR constraints.

The experimental results show that the proposed approach can reliably detect acoustic UAV emissions, but in a real scenario the false-positive rate would limit the usage of the raw classification output. Considering the recommendation of multi-modal detection systems in [20], it may become viable to combine multiple acoustic sensor outputs in array processing schemes, or audio-visual information into the detection process. Since some of the applications of drone detection may feature interdiction strategies, the inclusion of a low-complexity classification approach in a multi-object-tracking framework as described in [21] could prove vital to ensuring the safety of locations easily disrupted such as airports, or public events.

References

[1] Zeeshan Kaleem and Mubashir Husain Rehmani. "Amateur drone monitoring: State-of-the-art architectures, key enabling technologies, and future research directions". In: *IEEE Wireless Communications* 25.2 (2018), pp. 150–159.

[2] Rick L Sturdivant and Edwin KP Chong. "Systems engineering baseline concept of a multispectral drone detection solution for airports". In: *IEEE Access* 5 (2017), pp. 7123–7138.

[3] Sara Al-Emadi et al. "Audio Based Drone Detection and Identification using Deep Learning". In: *2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC)*. IEEE. 2019, pp. 459–464.

[4] Muhammad Zohaib Anwar, Zeeshan Kaleem, and Abbas Jamalipour. "Machine learning inspired sound-based amateur drone detection for public safety applications". In: *IEEE Transactions on Vehicular Technology* 68.3 (2019), pp. 2526–2534.

[5] Andrew W Christian and Randolph Cabell. "Initial investigation into the psychoacoustic properties of small unmanned aerial system noise". In: *23rd AIAA/CEAS Aeroacoustics Conference*. 2017, p. 4051.

[6] József Mezei, Viktor Fiaska, and András Molnár. "Drone sound detection". In: *2015 16th IEEE International Symposium on Computational Intelligence and Informatics (CINTI)*. IEEE. 2015, pp. 333–338.

[7] József Mezei and András Molnár. "Drone sound detection by correlation". In: *2016 IEEE 11th International Symposium on Applied Computational Intelligence and Informatics (SACI)*. IEEE. 2016, pp. 509–518.

[8] Sungho Jeon et al. "Empirical study of drone sound detection in real-life environment with deep neural networks". In: *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE. 2017, pp. 1858–1862.

[9] Jens Schroder et al. "Classifier architectures for acoustic scenes and events: implications for DNNs, TDNNs, and perceptual features from DCASE 2016". In: *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)* 25.6 (2017), pp. 1304–1314.

[10] Yoojeong Seo et al. "UAV Detection Using the Cepstral Feature with Logistic Regression". In: *2018 Tenth International Conference on Ubiquitous and Future Networks (ICUFN)*. IEEE. 2018, pp. 219–222.

[11] Yoojeong Seo, Beomhui Jang, and Sungbin Im. "Drone Detection Using Convolutional Neural Networks with Acoustic STFT Features". In: *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE. 2018, pp. 1–6.

[12] Andrea Bernardini et al. "Drone detection by acoustic signature identification". In: *Electronic Imaging* 2017.10 (2017), pp. 60–64.

[13] Dan-Ning Jiang et al. "Music type classification by spectral contrast feature". In: *Proceedings. IEEE International Conference on Multimedia and Expo*. Vol. 1. IEEE. 2002, pp. 113–116.

[14] Shlomo Dubnov. "Generalization of spectral flatness measure for non-gaussian linear processes". In: *IEEE Signal Processing Letters* 11.8 (2004), pp. 698–701.

[15] Sebastian Böck and Gerhard Widmer. "Maximum filter vibrato suppression for onset detection". In: *Proc. of the 16th Int. Conf. on Digital Audio Effects (DAFx). Maynooth, Ireland (Sept 2013)*. Vol. 7. 2013.

[16] Brian McFee et al. "librosa: Audio and music signal analysis in python". In: *Proceedings of the 14th python in science conference*. Vol. 8. 2015.

[17] Adam Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach et al. Curran Associates, Inc., 2019, pp. 8024–8035.

[18] Isabelle Guyon et al. "Gene selection for cancer classification using support vector machines". In: *Machine learning* 46.1-3 (2002), pp. 389–422.

[19] Max Kuhn et al. "Building predictive models in R using the caret package". In: *Journal of statistical software* 28.5 (2008), pp. 1–26.

[20] Gabriel Carisle Birch, John Clark Griffin, and Matthew Kelly Erdman. *Uas detection classification and neutralization: Market survey 2015*. Tech. rep. Sandia National Lab.(SNL-NM), Albuquerque, NM (United States), 2015.

[21] François Grondin and François Michaud. "Lightweight and optimized sound source localization and tracking methods for open and closed microphone array configurations". In: *Robotics and Autonomous Systems* 113 (2019), pp. 63–80.