Experimenting with Professional Microphones to Apply Acoustic Event Detection to Unmanned Aerial Vehicles

Kevin Hock[1], Mario Seideneck[1], Christoph Sladeczek[1], Michael Taenzer[1]

[1] *Fraunhofer Institute for Digital Media Technology IDMT, 98693 Ilmenau, Germany, kevin.hock@idmt.fraunhofer.de*

Introduction

Unmanned aerial vehicles (UAVs) are already in use for a wide variety of tasks for which they are equipped with various sensors such as video and thermal imaging cameras or flight assistance systems. Acoustic sensors on the other hand have not yet been widely adopted for UAVs, even though they offer a wide range of possible applications in combination with AI-based signal processing. Acoustic event detection (AED), e.g., could extend the sensing capabilities of UAVs by enabling them to react to user-defined acoustic events automatically. Possible scenarios include locating emergency situations, autonomous flight to the event location, automatic monitoring of the acoustic environment, or triggering an alarm. The acoustic sensor system required for AED must be robust to external environmental conditions as it is inevitably subject to perturbations such as the air flows through the rotors, operational noise, or wind noise. In this work, we have investigated a set of commercially available microphones for their suitability for such a system. The H520 hexacopter by Yuneec International Co. Ltd. was used as the carrier. We applied linear support vector machines (SVMs) in conjunction with OpenL3 embeddings for the realization of AED. To train the system, a dataset from the DCASE2020 challenge has been extended with recorded noise to consider different use cases.

Acoustical behavior of the drone

Acoustic sensors attached to drones are exposed to various disturbing noises, such as the presence of wind. Additionally, inherent noise also plays a significant role. To determine a suitable position for the sensor, the acoustic behavior of the drone needs to be considered. In order to analyze the sound radiation, the hexacopter was installed in an anechoic chamber of accuracy class 1 at Fraunhofer IDMT [1]. The measuring setup is visualized in Figure 1.

The drone was positioned at the center of a circular microphone array with a radius of 1 m using 1/4 " Microtech Gefell M360 microphones [2]. The measurements were then performed with a duration of 10 s at a sampling rate of 96 kHz. The H520 was oriented so that a boom axis with two opposing drone motors was in the same plane as the circular array. As the objective provides for recording sound sources on the ground, the microphone distribution of the array was oriented with the center of the hemisphere below the drone. Consequently, the microphone distribution on the array was non-equidistant, with the highest density of approx. 5.5° in the angular range from 157.5° to 202.5°. The reference direction of 0° corresponded to the position vertically above the H520. In this area, there was only one measuring microphone every 22.5°. Furthermore, three additional microphones were placed under the drone body at a radius of 25 cm, and one more in the center at a distance of 10 cm. Using these additional microphones, the influences of the distance to the drone were analyzed.

Figure 1(c) shows the polar plot of the determined root mean square (RMS) levels of the drone. In the area between microphones 9 and 17, the noise generated by the airflow of the rotors is clearly visible. By analyzing microphone 28 it becomes obvious that closer positioning to the body of the drone decreases the RMS level. Similarly at the rotor plane (microphones 5 and 21), a low RMS level can be observed. Regarding the detection of sound sources on the ground, these positions are not suitable for mounting microphones.

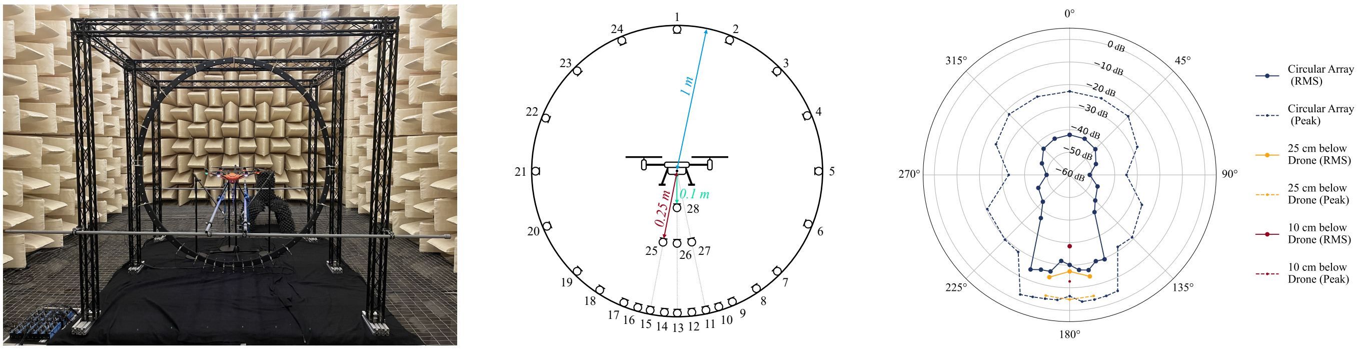
Besides the RMS levels, the peak levels are also shown in Figure 1(c). The significant discrepancies between the RMS and peak levels are due to slightly different rotor speeds caused by the drone's automatic flight control. These speed variations cannot be addressed within the scope of this investigation due to the lack of interfaces.

From the measurement results it can be summarized that an acoustic sensor should be mounted at a low distance below the drone's body. This way the shadowing effect of the drone itself can be utilized to reduce interference effects on the microphone. Furthermore, this allows an unaffected alignment of the directional characteristic to the ground.

Microphone selection and positioning

The objective of this study was to investigate the suitability of professional microphones for realizing drone supported AED. Four Sennheiser microphones with different directional characteristics were used for this purpose [3]. In order to assess the effect of the directional characteristics on the reduction of noise interference, the various microphones were positioned one next to the other at a distance of 15 cm as well as 25 cm below the drone. At each measuring position, the measurement was performed both with and without the provided foam windscreen. The microphones were calibrated via the substitution method, using a speaker at a distance of 1.5 m emitting a 1 kHz sine tone [4].

Table 1 lists the measurement results. The MKH 8040 and MKH 8050 microphones have the lowest attenuation. This can be explained by their cardioid and supercardioid characteristics. Even though an ideal cardioid pattern



(a) Mounting system for drone measurements in anechoic chamber.

(b) Measurement array setup.

(c) Polar plot of the noise emitted by the drone.

Figure 1: Measuring setup and result to determine the noise radiated by the Yuneec H520 drone.

Mic	RMS-Level [dB]			
	No wind protection		Wind protection	
	15 cm	25 cm	15 cm	25 cm
MKH 8040	-16.3	-16.7	-29.6	-29.7
MKH 8050	-17.6	-18.3	-28.8	-30.1
MKH 8060	-32.4	-30.2	-39.9	-38.3
MKH 8070	-47.8	-45.9	-54.0	-46.7

Table 1: RMS levels of the individual microphones. Highlighted cell marks the best result.

results in cancellation at rear sound incidence, the attenuation at 90° is only -6 dB, which leads the microphone to pick up large portions of drone noise and wind effects. A supercardioid microphone achieves a theoretical attenuation of -8.6 dB at 90° sound incidence. However, an attenuated side lobe from the 180° direction is disadvantageous. Consequently, rear incident sound components are recorded. These attenuation values are idealistic, and real microphones deviate from them for physical reasons. By contrast, an interference tube microphone is characterized by its high directivity, depending on the tube length. Using the MKH 8070, the attenuation is higher than with the MKH 8060, since the directivity of the interference tube begins at lower frequencies. Further attenuation of the drone interference could be achieved provided that the windscreen of the microphone was still within the shadowing area by the drone body. This was not valid for the MKH 8070 at a distance of 25 cm below the drone because of the size of the microphone. Therefore, the MKH 8070 is mounted 15 cm below the drone body in the prototypical evaluation.

Acoustic event detection

This feasibility study focused on a general investigation of implementing AED with a drone. No specific use cases were investigated, but rather the general detection of specific signals.

Datasets

The DCASE2020 Challenge (Task 5: Urban Sound Tagging With Spatiotemporal Context) datasets were used as baseline data. It comprises ten-second recordings of various acoustic sensors from New York City, and is divided into eight main classes: *Engine*, *Machinery Impact*, *Non-Machinery Impact*, *Powered Saw*, *Alert Signal*, *Human Voice*, *Music*, and *Dog* [5]. With regard to the data acquisition procedure of mentioned DCASE2020 dataset, the individual audio files consist of various combinations of these event classes. Such combinations are also referred to as soundscapes [6]. To preserve the original percentage distribution of the dataset after splitting it into all of the subsets (train, validation and test), any combinations occurring less than six times were augmented by pitch shifting and time stretching.

The initial dataset primarily consists of urban sounds, but many application scenarios of drones of this type take place in rural areas. To account for this, we additionally used Fraunhofer IDMT audio recordings of the rainforest of the *Taï National Park*, Côte d'Ivoire [7], and partly mixed the initial dataset with the forest recordings in an SNR interval from -30 LUFS to -6 LUFS using the Python library *Scaper* [6]. To consider these additional sounds within the annotations, the class *Non-Machinery Impact* - comprising general environmental sounds - was added to the extended soundscapes. Any newly introduced underrepresented class combinations were removed. Finally, this extended dataset consisted of 19,791 soundscapes, and we denote it as DR. In practice, an acoustic drone sensor would also be exposed to airflow as well as interfering sounds caused by the drone itself. Furthermore, the directional attenuation of environmental noise by microphone characteristics is limited. Therefore, an additional dataset was generated by combining DR with variable drone noise from the Yuneec H520 at different SNR intervals, prior to removing underrepresented combinations. This dataset subsequently comprised 19,972 soundscapes, and is referred to as DRD. Both DR and DRD were split with a 66%-17%-17% ratio into train, validation and test subsets.

Feature extraction and classification method

The performance of a machine learning (ML) algorithm can be critically dependent on the size of the available dataset. According to the specific task, the creation of large datasets is very time consuming and costly. Various deep learning models were trained in a supervised or self-supervised fashion on large datasets in the field of acoustic event detection. These pre-trained models are designed to learn complex feature representations with strong discriminative power, called embeddings. Through transfer learning, the models can be used to extract these embeddings from smaller datasets. In this study, OpenL3 was used to extract the embeddings due to their good performance in combination with linear SVMs [8]. The following settings were used: content type environmental videos, mel-spectrogram with 256 bands, and an embedding dimensionality of 512.

SVMs have originally been developed to perform binary classifications. Two classes are to be separated by a hyperplane in the feature space using as few data points as possible (the support vectors). The SVM algorithm chooses this hyperplane such that the distance between the class boundaries is maximized. For subsequent classifications, only the support vectors are needed [9, 10]. Since the fundamental SVM algorithm is not suitable for classification of larger numbers of classes, we split the detection task into several binary one-vs-rest classifications, where the idea is to separate one of the available classes from the remaining classes.

Experimental design and evaluation

A design criterion for our study was that the evaluation has to be performed wrt. a liberal classifier. For the monitoring task, a high recall is required to detect as many events as possible. Therefore, we settle on experiments with recall thresholds of $R = 0.7$ and $R = 0.8$ for the performance evaluation of the SVMs [11]. Within this range, the local maximum of the F_1-scores (Eq. 1) is determined on a class-by-class basis to derive the associated decision threshold for each class.

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN} \quad (1)$$

First, we consider the classifiers trained on DR. The classifiers show high F_1-scores. It is noticeable that classes occurring more frequently are better classified in contrast to the more underrepresented ones, for example in *Human Voice* vs. *Dog*. While the performance of *Engine*, *Non-Machinery Impact* and *Human Voice* remains constant with a minimum recall of 0.8, classification of the other classes deteriorates. This can be attributed to the fact that their best possible operating points, as measured by the maximum F_1, are below this recall threshold.

The focus is on the selection of a liberal classifier. The SVM with a regularization parameter of $C = 0.1$ was chosen for the further evaluation steps as it achieved the highest F_1-scores for both recall thresholds over most of the classes of the application scenarios. However, it should be noted that the classification thresholds of individual classes are sometimes significantly below 50%. This shows that the classifier assigns the positive class for rather underrepresented classes at lower probabilities, which inevitably worsens the precision.

With DRD, a significant degradation of the classifications becomes apparent. Here, the lower the SNR between drone sounds and the original soundscapes, the more the performance is degraded. Even for these SVMs, varying the C-parameter leads to no or only minor changes in the metrics.

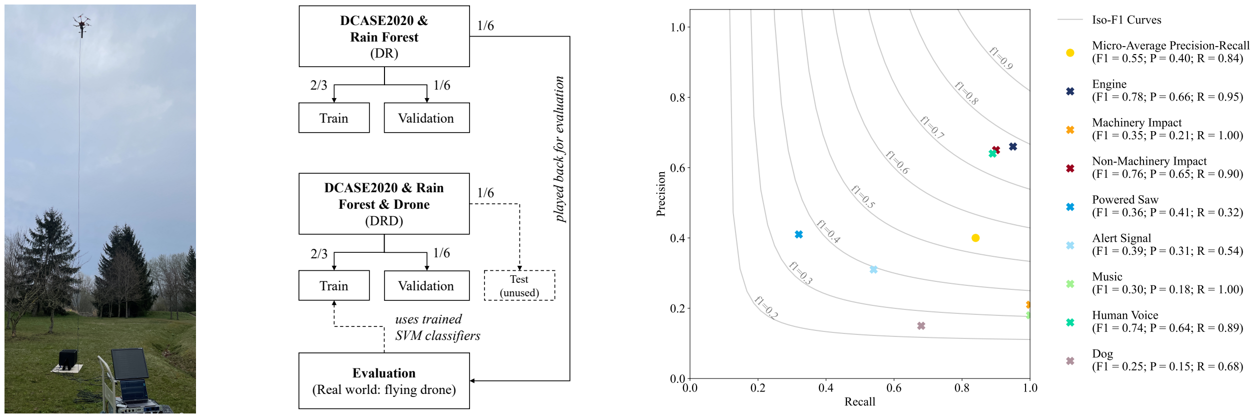
In summary, these results show that the combination of OpenL3 embeddings and linear SVM can potentially lead to a performant detection of relevant classes under ideal conditions. However, drone noise in the microphone recordings is inevitable in a real in-flight scenario. Therefore, a classifier needs to be able to handle such noises without significantly degrading classification performance. Further investigations are needed.

Real-world evaluation

When using a drone system to detect acoustic events in real-world conditions, numerous uncontrollable interfering factors occur, such as wind or ambient noise. According to the results of the microphone investigations, a prototype system was built using the MKH 8070 with the Yuneec H520. The microphone was prepared with its windscreen and mounted freely movable 15 cm below the drone. In order to have a defined sound source, two Seeburg TSNano speaker were placed on the ground and oriented to radiate upwards. The sound pressure level was calibrated using a white noise signal to achieve 100 dBSPL at a distance of 1 m above the speakers. The drone flights were conducted at a height of 10 m above the ground on a meadow near the Fraunhofer IDMT. The setup is depicted in Figure 2(a). According to the weather station of the Ilmenau University of Technology, the wind speed on the ground was approx. 0.1 m/s. However, light gusts occurred during test execution. Due to the limited flight time of the multicopter, only 150 soundscapes ($< 1\%$) of the evaluation data were used.

Figure 2c shows the classification results of the test flight. The *Engine*, *Non-Machinery Impact* and *Human Voice* classes achieve a recall between 0.89 and 0.95, and a F_1-score of 0.74 to 0.78. This implies that in terms of a liberal classifier, a very good classification is possible. Although *Machinery Impact* and *Music* show a recall of 1.0, the F_1-scores are only around 0.3 due to a very low precision of ≤ 0.21. Consequently, the number of false positives clearly predominates here. In the case of the remaining classes, F_1-scores of < 0.4 also indicate insufficient performance.

In summary, the classes most frequently represented in the train set tend to show a practical classification in the application. However, these results may have been negatively impacted by a number of issues. One of them is potential misannotations in the original DCASE2020 dataset. In addition, due to the takeoff mass of the drone and prevailing wind at an altitude of 10 m, positioning the multicopter precisely above the speakers proved difficult. This may negatively influence the results due to the high directivity of the microphone, but at the same time this highlights the problems in a real-world scenario. More extensive test flights using the full evaluation set would be required for a final assessment.



(a) Flight of the prototype system.

(b) Experimental design.

(c) Classification results.

Figure 2: Evaluation results of the prototype system consisting of a Yuneec H520 multicopter equipped with a Sennheiser MKH 8070. The system performed classification tasks of acoustic events under real-world conditions using two speakers as the sound source. P = Precision, R = Recall.

Conclusions

The evaluation of the prototypical implementation was done by automated playback and parallel recording of selected soundscapes. These recordings were submitted to feature extraction and finally to classification. The results under controlled conditions indicate that the classifiers can exhibit high recall and F_1-score on the *Engine*, *Non-Machinery Impact*, and *Human Voice* classes. The best results were obtained when drone noise with an SNR of -6 to 10 LUFS is included in the training. From this proof of concept, it can be concluded that AED in conjunction with a multicopter is generally feasible. Nevertheless, these are only tendencies which must be further investigated by optimizing the individual components of the concept, and supplementary evaluation steps.

Acknowledgements

The authors like to thank Sennheiser electronic GmbH for providing microphones to support the investigations.

References

[1] DIN EN ISO 3745:12, “Acoustics - Determination of sound power levels and sound energy levels of noise sources using sound pressure - precision methods for anechoic rooms and hemi-anechoic rooms,” DIN e. V., Tech. Rep., Oct. 2012.

[2] IEC 61672-2, “Electroacoustics - Sound level meters - Part 2: Pattern evaluation tests,” International Electrotechnical Commission IEC, Tech. Rep., Sep. 2013.

[3] Sennheiser electronic GmbH & Co. KG, “Mikrofone,” https://de-de.sennheiser.com/mikrofone?order_by=relevance&page=3, Tech. Rep., 2021-06-01.

[4] DIN EN IEC 60268-4:2019-08, “Sound system equipment - Part 4: Microphones,” DIN e. V., Tech. Rep., Aug. 2018.

[5] J. P. Bello, C. Silva, O. Nov, R. L. Dubois, A. Arora, J. Salamon, C. Mydlarz, and H. Doraiswamy, “SONYC: A system for monitoring, analyzing, and mitigating urban noise pollution,” *Communications of the ACM*, vol. 62, no. 2, pp. 68–77, 02 2019.

[6] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, “Scaper: A library for soundscape synthesis and augmentation,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2017, pp. 344–348.

[7] S. Heinicke, A. K. Kalan, O. J. Wagner, R. Mundry, H. Lukashevich, and H. S. Kühl, “Assessing the performance of a semi-automated acoustic monitoring system for primates,” *Methods in Ecology and Evolution*, vol. 6, no. 7, pp. 753–763, 2015.

[8] J. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, “Look, listen, and learn more: Design choices for deep audio embeddings,” in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3852–3856.

[9] A. V. Joshi, *Machine Learning and Artificial Intelligence*, 1st ed., ser. Springer eBook Collection. Springer, 2020.

[10] R. F. de Mello, *Machine Learning. A Practical Approach on the Statistical Learning Theory*, ser. Springer eBook Collection. Springer International Publishing, 2018.

[11] A. Mesaros, T. Heittola, and T. Virtanen, “Metrics for polyphonic sound event detection,” *Applied Sciences*, vol. 6, no. 6, 05 2016.