UAV identification from acoustic signals using statistical learning : a state-of-the-art

Antoine Purier, Simon Bouley, Lucille Pinel-Lamotte

QUIETDRONES 2024
MANCHESTER
8th - 11th September 2024

UAV identification from acoustic signals using statistical learning : a state-of-the-art

Session: Acoustic Detection and Identification of Drones

Antoine Purier, MicrodB, France, antoine.purier@microdb.fr
Simon Bouley, MicrodB, France, simon.bouley@microdb.fr
Lucille Pinel-Lamotte, Vibratec, France, lucille.pinel@vibratec.fr

Abstract Over the last few years, the deployment of Unmanned Aerial Vehicles (UAVs) has sky-rocketed, pushed by the ever-growing range of commercial applications and various malicious purposes. In order to protect sensitive facilities and public areas, their effective and prompt localization and identification is necessary. As no modality (electro-optical, radio, etc.) is yet able to single-handedly perform both of these tasks, a combination of several sensors is required to fulfill these objectives under all real life conditions. Among the different modalities, acoustics relies on pressure signals captured by single microphones or phased arrays to carry out these tasks and has proven its detection efficiency for short distances. Whereas the localization step can be achieved by a set of well-known techniques widely described in the literature, the identification stage has not reached the same level of maturity. Therefore, this paper focuses on the latter and aims at presenting a state-of-the-art of classification techniques based on statistical learning and dedicated to UAV acoustic signatures. Additionally, a specific care is taken to assess the suitability of the reviewed techniques to real life conditions and requirements.

Keywords: Acoustic localization, Drone identification, Machine Learning, Deep Learning

1. Introduction

Recent developments on the geopolitical scene have shown the increasing use of UAVs as a low cost and effective vector of defense and attack, whether it is as part of global conflicts or regional terrorist actions. The use of UAVs for local industrial spying is also of great concern. Many solutions have emerged to detect UAVs, based on different modalities. To this day, no solution has proven efficient enough to cover this detection task single-handedly in all scenarios. Electro-optical solutions, although mature and very efficient under good weather conditions, will struggle during dark cloudy nights and rainy episodes. Solutions detecting radio-frequency communications between pilots and UAVs are soon to prove pointless with the emergence of AI-piloted UAVs. Radar detection struggles at detecting small objects and is impaired under heavy rain. Moreover, cardboard UAVs have been employed in Ukraine, as cheap and undetectable alternatives. Even though industrials have made considerable efforts over the last few years to reduce UAVs acoustic footprint, to meet market regulations

and get accepted by customers, it is a much harder task to make UAVs completely silent than to make them invisible to radars or cameras. Taking advantage of progresses made in the statistical learning field, many works addressing UAVs acoustic-based detection have been published to date. Traditional machine learning algorithms have been tried over and over and have proven relatively efficient in controlled situations. More advanced and recent deep learning methods have been implemented with success over the last 10 years, boosted by the rise of similar architectures used in computer vision and natural language processing applications. Overall, most works have applied their efforts on mono unprocessed signals recorded in relatively quiet environments and only few of them have leveraged array processing techniques in a joint effort for localization, denoising and identification in complex, real-world scenarios. Array processing techniques could thus play a key role in the preprocessing of algorithms input for improved detection performances in noisy environments.

This contribution starts by presenting a global procedure for joint acoustic localization and identification of UAVs, listing existing array signals processing techniques and how they integrate with statistical learning algorithms. It then reviews the latest works found in the literature on UAVs identification and related promising contributions on sound classification in general.

2. Global Procedure for Joint Acoustic Localization and Identification of UAVs

2.1 Array processing for source localization

UAVs identification in intricate and noisy environments generally involves two main stages denoted as localization and classification. Widely employed to recover the position of acoustic sources in space and time, phased-array techniques also act as signal gain amplifier. As these algorithms perform localization from multiple microphone signals acquired coherently, they also alleviate incoherent noise. Therefore, once the target source is correctly tracked, the gain amplifier provided by a microphone array help the extracted signals to emerge from background noise. Approximately equal to $10\log_{10}(M)$ (dB), with M the number of array microphones, this gain accentuates the need of acoustic imaging as a preprocessing step before any UAV classification method from acoustic signals.

High array performances (resolution of the main lobe of directivity and mainlobe-to-sidelobe ratio, as denoted as dynamics) are necessary to distinguish multiple sources in space and extract the clearest signals of each acoustic event from measurement. Indeed, array with a low number of microphones will generate artifacts that can hide sources of weak amplitude or mitigate close sources. Therefore, a specific care must be taken to localize sources such as UAVs when they operate in complex environments (with a high background noise or in presence of multiple sound events), as their acoustic radiation can often hardly emerge in the measurement. Consequently, even with microphone arrays, the extracted signals to identify are inevitably polluted by background noise, close sources, with an altered frequency spectrum.

Multiple signal processing techniques have been employed by authors to localize UAVs (Lamotte et al., 2020), but all methods do not provide the same outputs to feed statistical learning algorithms. Goniometry or TDOA (Time Direction Of Arrival) computes the generalized cross-correlation function between pairs of microphones (Knapp & Carter, 1976) to locate the origin of incoming waves. A set of four sensors is enough to localize sound sources in space and time, at the price of low resolution. Its application to UAVs detection remains restricted to situations with a limited number of sources and a high signal-to-noise

ratio (SNR) (Blanchard et al., 2019). Also, signals directly measured by the microphones are used as basis for the calculation of the identification descriptors, precluding the array to spatially filter the source signal. Direction of arrival can also be detected by the estimation of pressure and particle velocity at the center of arrays (Ramamonjy et al., 2018).

A second class of techniques includes high-resolution methods, such as MUSIC (MUltiple SIgnal Classification (Schmidt, 1986)), which achieves an eigenvalue decomposition of the measured signals to estimate signal and noise subspaces. Projecting the data onto the inverse of the noise subspace allows to recover the direction of arrival. This method obtains accurate localization performances (Baron et al., 2020) but needs a very high SNR to remain competitive. Also, the separation of signal and noise subspaces requires to know precisely the total number of sources to track, which is hardly achievable in practice.

Finally, a large amount of authors used beamforming-like methods to track UAVs in time domain (Blanchard et al., 2020), or in frequency domain (Herold et al., 2020). While Delay-and-Sum (DAS) beamformer suits well to track moving sources, Conventional Beamforming (CBF) (Chiariotti et al., 2019) must be applied as a pseudo-temporal processing to follow non-stationary sources. To do so, measured signals are cut into sequences of short duration for which frequency-based beamforming is performed (Baron et al., 2019). This technique allows to retrieve the position of the sources for each snapshot, but an additional step is needed to reconstruct the temporal signal of the tracked UAVs, by focusing along time the microphone signals in the direction of the CBF maps maxima. Deconvolution or high-resolution methods can be used to refine the resolution of the targeted sources, but the extracted time signals remain prone to DAS performances in terms of resolution and dynamics.

2.2 Localization outputs

As detailed before, various methods can be carried out to track UAVs, but all methods do not provide the same outputs. TDOA methods are well-suited to track single sources as almost all the coherent acoustic fields must be coming from a clear direction. With a high SNR, the time signal captured by one microphone remains sufficient to feed the identification process. With lower SNR, the gain amplifier provided by the spatial filtering of DAS beamforming allows to directly track the time signal of the source and alleviate background noise. If no spatial filtering is achieved, additional source separation and denoising processes must be carried out before the identification step to classify the different elements present in the acoustic scene. These two kinds of techniques can natively provide one time signal, related to one acoustic source, to classifiers. From time signals, one can define many descriptors in temporal, spectral and cepstral domains (Childers et al., 1977), often used in machine learning algorithms. Spectrograms can also be computed from time signals to draw frequency-time domain representation of the acoustic scene, or of the filtered tracked source. Considered as images, spectrograms are adapted to feed deep learning algorithms, massively used for image recognition.

Frequency-based localization algorithms were designed for stationary sources. If the speed of UAVs is not too high, it is reasonable to perform pseudo-temporal localization, cutting microphone signals into short sequences. For each sequence, a cross-spectral matrix (CSM) can be computed to perform conventional beamforming or any other high-resolution method. Additionally, using a Welch periodogram (Welch, 1967) to estimate the CSM can help to drop out incoherent noise. These algorithms can localize acoustic sources, but are not able to reconstruct their time signal. One way would be to compute descriptors in fre-

quency and cepstral domains only, or to stack in one image all spectra along time, for nodes of the calculation grid that correspond to targeted sources. An alternative solution is to perform DAS beamforming only for these nodes to track the sources signals, or to add a time domain deconvolution step to increase the array performances (Cousson et al., 2019). With frequency-based solutions, a priori information about the targeted source spectrum can be taken into account to finely tune the localization map (Leiba et al., 2022).

In conclusion, the method used to capture valuable inputs for UAVs identification greatly depends on the complexity of the environment. Localizing UAVs does not guarantee the quality of the input provided to identification algorithms. Measured signals can reveal the presence of multiple sources of various frequency properties with a wide range of amplitudes, overwhelmed by strong background noise. Spatial and frequency filtering may become unavoidable in severe measurement conditions, calling for advanced acoustic imaging techniques.

3. Brief Overview of Identification Methods Applied to UAVs

When addressing sound classification, several methods can be leveraged, from well established machine learning algorithms that have been around for decades to more recent deep learning architectures that are not always easy to comprehend and implement. While the former are known to be highly explainable, be easier to implement and require less data and compute power to provide satisfactory performances, the latter are still seen as black boxes, requiring much effort and compute resources as a price for their high performances. As already mentioned, these two families of statistical models train on different types of inputs. Machine Learning models learn on tabular data while most state-of-the-art neural networks dedicated to sound classification take 2D-matrices or images as input (mel-spectrograms, mel-frequency cepstral coefficients (MFCC) maps or other 2D representations of a sound event). For machine learning models, training data consist in manually engineered acoustic indicators calculated on short time sequences of labelled sound events. A model will ingest a table of n rows and m columns, where n corresponds to the number of observations or time sequences and m to the number of calculated indicators. It is thus necessary to fuel machine learning models with domain knowledge to create the most relevant indicators. State-of-the-art neural network architectures include this feature engineering process by finding the relevant information in the 2D representations of sound events they are given.

4. Machine Learning

Many applications of traditional machine learning algorithms for UAVs detection can be found in the literature, and a significant part of them are fed by mono channel signals recorded from single microphones.

Among them, Support Vector Machines (SVM) have been widely used for sound classification applications. SVMs distinguish between two or more classes by finding the optimal hyperplane that maximizes the margin between the closest data points of opposite classes (also called support vectors of each class). Each training sample is projected in an n-dimensional space, n corresponding to the number of features. The hyperplane is a line in a 2-dimensions space and a plane in a 3-dimensions space. In the case of sound classification tasks, features are acoustic metrics or statistics computed on the input signal. (Baron, 2020) calculates a set of acoustic descriptors on both temporal, spectral and cepstral signals, such as signal's *skewness*, *zero-crossing rate* or *Shannon's entropy*. The nature of these descriptors will heavily depend on the classes to distinguish. SVMs are able to handle non linearly separable data by transforming the training data from their original

space to a higher-dimensional space in which they can be linearly separated. Known as the "kernel trick", polynomial, gaussian or sigmoïd kernels are the most widely used functions to perform this task. An SVM model can either be implemented in a multi-class setting (ex: drone vs. helicopter vs. plane), a binary setting (ex: drone vs. helicopter) or in a one-class setting (drone vs. all other possible classes). One-class SVM is an unsupervised algorithm that learns a decision function for novelty detection: classifying new data as similar to or different from the training set. In the case of UAVs detection, only drone sound samples are fed to the model for training. At inference time, the model is confronted to all the classes that constitute the sound environment at the sensor location. Each sound event is either classified as pertaining to the drone cluster/class or not.

(Bernardini et al., 2017) propose a multi-class SVM trained on single microphone recordings of drones, nature, crowd, train and street recordings. Multiple binary classifiers are trained in a one-vs-one strategy (ex: drone vs. nature, nature vs. train, etc.) and their results are combined. Features such as *temporal centroïd*, *zero crossing rate*, *spectral roll-off* or *mel-frequency coefficients* are calculated on short 20 ms time sequences. These features are then statistically aggregated on *mid term* 200 ms time sequences. The computed statistics constitute the audio descriptors fed to the model. The resulting precision of the drone recognition is 98,3%. Although promising, this method sets a fixed and limited number of classes. Moreover, little information is given on the diversity within each class, and the number of drone models included in the *drone* class. It is thus likely that such a model would not be really suited for a deployment in a real-life environment, where classes are often mixed together and their occurrence changing from location to location and consequently difficult to exhaustively anticipate.

(Baron, 2020) proposes a joint localization-identification method feeding array sensor measurements to the MUSIC high-resolution algorithm for sources localization. Focalized signals on each localized sources are then passed to a One-Class SVM model for inference. Features used are temporal, spectral and cepstral based, computed on 200 ms time sequences of focalized signals. Good results are obtained for the identification part, but the environment is quite controlled and the number of classes limited, making it possibly less robust to real-life environments with many more co-existing sources of different classes.

(Shi et al., 2018) apply a Hidden Markov Model to Mel-Frequency Cepstral Coefficients for UAVs detection. Two schemes of different numbers of MFCC for training, 24 and 36, are compared. Three training datasets are set up with an increasing number of sound types. Each training and test sets is comprised of five clusters corresponding to the sounds of drones, planes in airports, cars, birds, and rain, respectively. Three models of drones are represented in the *drone* class. HMMs are trained on each single cluster. The best performance is obtained with 36 MFCC for each of the training sets. Different levels of gaussian white noise are also added to vary the SNR and test the robustness of the method. The recognition rate is still up to 80% with a SNR of 5dB in the best configuration of features and train set. This approach also boasts promising results. But as for the multi-class SVM approach, the number of classes is fixed and limited. Moreover, adding gaussian white noise is likely not representative of sources signature mixes that could occur in real-life in the case of recordings with a single sensor. Finally, this method is significantly harder to implement and resources are scarcer on the subject.

(Blass & Graf, 2020) first propose both a localization and a detection method of UAVs, and later combine them as a joint method. The localization stage is supported by a 31-microphones array and consists in a DOA estimation algorithm using a Gaussian Mixture Model (GMM) based approach to track multiple sound sources. The detection stage consists in a binary classification (drone vs. non-drone). Support Vector Machines and Random Forests as well as Feed-Forward Neural Networks and Recurrent Neural Networks are

trained and tested for this stage. A custom database of 10 hours of audio data from nine different UAV types is used to train and test these models. Several noise contexts and flight stages are accounted for. The models are first trained on mono audio recorded from the reference microphone at the center of the array. A mix of different types of audio features is proposed, such as generic low-level descriptors, MFCC with modified frequency ranges and custom features derived from tracked spectral peaks. This set of features is calculated on 20-200 ms time sequences of signal. Additionally, temporal smoothing is applied and statistics (min, max, mean) are computed over consecutive time sequences. The Random Forest model is chosen for its satisfactory performance and low execution time. This model is then evaluated, using the same data and model setup, on beamforming signals (delay-and-sum outputs of the three most prominent sources) to test the robustness of the model when integrated after the localization stage. The F1-score metric is used to evaluate it. F1-score is a measure of the harmonic mean of precision and recall. A high F1-score generally indicates a well-balanced performance, demonstrating that the model can concurrently attain high precision and high recall. A F1-score of 0.85 is obtained on the test set of beamforming signals. This contribution is really promising has it efficiently managed to join both localization and detection methods. Prior localization of sources coupled with delay-and-sum permits to give clean inputs to the model, facilitating detection. However, no exhaustive information is given on the non-drone classes included in the train, validation and test sets. Also, a binary classification is expected to work fine for a fixed and limited number of non-drone classes but will likely not scale well if the sensor is deployed at multiple locations where new unseen classes may occur.

5. Deep Learning

Deep Learning has known a relatively recent rise, mainly driven by the availability of ever more data originating from sensors and smartphones and the drop in costs for computer resources. Initially boosted by computer vision applications such as autonomous driving or face recognition, Convolutional Neural Networks (CNN) have quickly become go-to architectures due to their extensive documentation and the high availability of pre-trained models. The sound classification field has successfully adopted such architectures by transforming time sequences modelling into an image classification problematic.

Recurrent Neural Networks (RNN), first conceptualized in the early 1980s, is another class of networks that can model sequences of arbitrary lengths, making them suitable for tasks such as natural language processing, time series forecasting or speech recognition. Such models maintain a memory of what has been calculated in previous elements of a sequence and use it for subsequent predictions. One of the main problems of RNN tough is its incapacity to handle sequences that are too long, for which case the retro-propagated error gradients can either become too big and explode or too small and vanish as they progress through the network. In such scenario, the model is incapable of distinguishing information to retain or not and to effectively learn. Several optimizations of RNN have been proposed to alleviate such limitations. Long Short-Term Memory (LSTM) network was proposed in (Hochreiter & Schmidhuber, 1997) to answer RNN's vanishing gradient and long-term dependency issues. It was continuously improved over the years. Gated Recurrent Unit (GRU) network was proposed in (Cho et al., 2014) as a variation of the LSTM network, introducing an update gate and a reset gate. With less parameters than an LSTM unit, the GRU is more flexible and takes less time to train but at the cost of slightly poorer performances.

Convolutional Recurrent Neural Networks (CRNN) have been created by mixing both a CNN and an RNN. In such configurations, the convolution and pooling layers of a CNN architecture extract features that are then fed to the GRUs or LSTM units to capture temporal information that a CNN only would struggle to get.

CNNs, RNNs or combined architectures are the most used architectures to date for UAVs identification.

In (Alla et al., 2024), an LSTM and a CRNN model are trained on a dataset comprising 130 10-seconds long mono channel audio pertaining to four classes: *airplane*, *bird*, *drone*, *helicopter*. Both models are trained first on mel-spectrograms and then on MFCC maps for comparison. Each model is trained in a binary classification setting (3 non-drone classes vs. drone class) and in a multi-class setting (each class individually). The CRNN model outperforms the LSTM model in all configurations of features and classes grouping, and takes significantly less time to train. Drone and helicopter WAV recordings used for training are issued from the dataset detailed in (Svanström et al., 2021). This dataset is created, for the audio part, from outdoor measurements in airports environments. Three drone models of different weights constitute the *drone* class. Audio are recorded with a Boya BY-MM1 mini cardioid directional microphone.

(Utebayeva et al., 2022) propose a real-time recognition framework focusing on the detection of loaded UAVs. A simple RNN, an LSTM, a Bidirectionnal LSTM (Schuster & Paliwal, 1997) as well as a GRU are trained and their performances compared on the task of classifying sound samples under three classes: unloaded UAVs, loaded UAVs and background noises. Multiple models of drones, loaded with varying payloads, were recorded in an open environment using a single microphone. Diverse background noises were collected during the measurement campaign. Some UAV sounds from open sources were also used as part of the data collection process. Mel-spectrograms were computed on 1s sound samples and fed to the networks. The GRU architecture with 64 cells was found to have a high degree of predictability, being able to recognize loaded UAVs and unloaded UAVs with 98% accuracy and background noise with 99% accuracy. This work proposed a successful solution for drone detection on relatively short durations of audio signals. However, it does not include a localization stage and only relies on signals captured with a single microphone. It is thus likely that it could struggle in more challenging environments. No detail is given on the ratio of noisy UAV sound samples or on the actual performances of investigated models on UAV sounds overlapping with other sources.

A joint localization and identification deep learning approach based on raw microphone arrays temporal signals has been proposed in (Bavu et al., 2022). Based on the *BeamLearning-ID* convolution based model introduced in (Pujol et al., 2021), this approach, as opposed to those based on traditional 2D input CNN models, projects the temporal data into the most appropriate representational space, eliminating the a priori behind the computation of a 2D spectral representation as input. The model leverages atrous 1D convolution filter banks dividing into two branches, one for drone recognition and one for real-time 3D localization. This “Joint Feature Learning” approach also has the advantage of minimizing inference latencies, which is a key criteria in defense applications. The reported 3D absolute angular localization error has a median of less than 4° and the rate of non-detection of a drone is of 1 %. The true-class inference rate, the performance at identifying drones models, is of 78 %. However promising, this approach, by not processing temporal microphone array signals to reconstruct sources through deconvolution methods, seems only limited to the localization and detection of one source only.

Another joint localization and identification method was proposed in (Sun et al., 2023). The proposed approach utilizes a 6-microphones array with a beamforming algorithm to localize UAVs and capture their sound signature filtered from unwanted interference and noises. It then computes both mel-spectrograms and MFCC maps from the audio signal for comparison, which are subsequently input to a CNN for classification. Each sound sample used for training the model is segmented into 128 sequences of 23 ms, overlapping by 50 %, creating 128 x 128 inputs to the model. This work concludes in a greater detection

accuracy when using MFCC features instead of mel-spectrograms. While successful and promising, this contribution only evaluates UAVs hovering at fixed locations, thus being able to reconstruct feature maps over long durations (over 1s). In real scenario however, it is likely not to be the case as UAVs are moving fast and many other sources are likely to be present in the scene, leading to shorter available time samples for inference.

Introduced in (Vaswani et al., 2017), the Transformer architecture has been a major recent innovation in the world of Deep Learning. Originally motivated by Natural Language Processing (NLP) applications, this new architecture is based solely on attention mechanisms, dispensing with recurrence and convolutions entirely to process sequential inputs. This type of architecture helps better capture long-range global context, and offers an alternative to Recurrent Neural Networks and their vanishing gradient limitations. (Dosovitskiy et al., 2020) proposed one of the first applications of such an architecture to image classification and showed the excellent performances obtained by its Vision Transformer (ViT) compared to state-of-the-art convolutional networks, while requiring fewer computational resources to train. The Transformer, applied directly to sequences of image patches, is pre-trained on large amounts of data and transferred to multiple mid-sized or small image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etC.). Following this effort, (Gong et al., 2021) proposed the Audio Spectrogram Transformer (AST), a direct application of the Vision Transformer to sound, where 2D audio spectrograms inputs are split into sequences of 16x16 patches with overlap and then linearly projected to sequences of 1D patch embeddings. Each patch embedding is added with a learnable positional embedding and an additional classification token is prepended to the sequence. The output embedding is passed as input to the Transformer whose output classification token is used for classification with a linear layer. Pre-trained on real-world images and fine-tuned on Google's AudioSet dataset comprising 2 million 10-seconds 527 classes audio clips, AST achieves new state-of-the-art performances when fine-tuning on substantially smaller datasets (95.6% accuracy obtained on ESC-50, a 2000 5s 50 classes environmental sounds dataset). Further improvements have been achieved in (Koutini et al., 2021) which proposes the PaSST architecture. Based on AST, it introduces *Patchout*, a technique that drops parts of the Transformer's input sequence when training, forcing the Transformer to perform the classification using an incomplete sequence. This technique also substantially speeds up training time. Performance close to state-of-the-art is achieved when fine-tuning on the ESC-50 dataset. All variants of the PaSST model can be fine-tuned on this dataset in less than 5 minutes on a single consumer-grade GPU.

Application of the Transformer architecture to UAVs identification has only been found in (Anidjar et al., 2023). In this work, a framework to detect mechanical anomalies in sound emitted by UAVs is presented. This framework uses a Transformer based model for embeddings extractions followed by a VGG network, a classical deep CNN architecture introduced in (Simonyan & Zisserman, 2014). The Transformer is not used in this paper as a standalone classification model.

6. Discussion

As presented in this literature review, a various range of techniques has been deployed to successfully locate, track and identify UAVs. Phase-array techniques are able to follow moving sources whose acoustic signatures can feed statistical learning algorithms to classify them as UAVs or not. However, a closer study shows that a joint localization-identification method, highly reliable and robust to heavy environmental conditions is still to develop. To the best of our knowledge, no study has proved the efficiency of a specific method to track and identify multiple drones in presence of a strong background noise and disturbing sources. It also appears that the performances of any statistical learning highly depends

on the quality of the input data, both for the learning and testing steps. In most situations, real-life measurements of UAVs with single microphones cannot be directly used as input data, as the acoustic fingerprint does not emerge enough from background noise. This way, the similarity between UAVs signals used beforehand to train the identification model and measured signals to test becomes too weak, leading to misestimation. Therefore, the amplitude gain brought by the addition of microphones in phased-array techniques is needed to make the signal of interest emerge. However, the output of acoustic imaging techniques is no longer a raw microphone signal, but some extracted data (in time or in frequency domain) from an acoustic map, more or less filtered by the method employed. According to the acoustic scene where the UAVs are tracked, a set of requirements can be considered. First, some acoustic imaging methods need to know the number of sources to track to be efficient (MUSIC) or can struggle to follow multiple sources of different nature (TDOA). Second, a localization technique with a high robustness to low SNR environment is a prerequisite to most real-case situations. On one hand, frequency-domain phased-array techniques have proved their capacity to drop out incoherent noise or to filter some specific frequency bands to make the signal of interest emerge. However, in case of moving sources, the time sequences can be drastically shortened, unsettling the good convergence of the signal averaging. Also, unless only frequency features are used for the statistical learning, the time signal of the localized sources must be reconstructed all along the drones trajectories. One the other hand, time-domain acoustic imaging techniques can directly extract signals of interest to compute time features or spectograms, but are prone to low array performances (resolution and dynamics) unless preprocessing steps such as frequency filtering (with some a priori information) are performed beforehand. With such low performances, sources of interest can be hidden by disturbances sources and their acoustic signatures polluted by unwanted sources and recognized as such. Finally, a trade-off is to be found between the accuracy of the source signal reconstruction and the computational burden of the algorithm to be able to follow and identify moving UAVs in real-time conditions.

It appears clearly that a great deal of efforts has been made to improve classification and identification algorithms dedicated to acoustic signals. However, the way these signals are measured are still prone to misestimation, due to a large set of bias. This review shows that the identification of a single UAV in a quiet and controlled area, close to the microphone array is achievable by several techniques. However, if studies become to tackle the tracking of multiple UAVs (Herold et al., 2020) or the localization of drones in low SNR environments (Wu et al., 2024), new ideas are still needed to reliably identify UAVs in real-life conditions with the acoustic modality.

7. Conclusion

UAVs identification has recently become an even more crucial field of research. While former contributions relied on machine learning algorithms to perform such task, deep learning methods seem to have gained popularity and most recent works study their effectiveness. Although many publications present excellent performances, regardless of the statistical learning methods used, a lot of them remain far from real-case scenario. Joint localization and identification methods are the most promising ones, for their ability to still perform in high background noise environments and output the position of detected UAVs. Still, progresses need to be made to produce such robust solutions that work in a variety of environments, on very short segments of audio, also tackling inference times problematics. With the advent of foundational deep learning models pre-trained on huge datasets, fine-tuning deep learning models on downstream tasks with substantially less data available is now easily feasible and remains to be investigated for the task of UAVs detection.

REFERENCES

Alla, I., Olou, H. B., Loscri, V., & Levorato, M. (2024). From sound to sight: Audio-visual fusion and deep learning for drone detection. In *Proceedings of the 17th acm conference on security and privacy in wireless and mobile networks* (pp. 123–133).

Anidjar, O. H., Barak, A., Ben-Moshe, B., Hagai, E., & Tuvyahu, S. (2023). A stethoscope for drones: Transformers-based methods for uavs acoustic anomaly detection. *IEEE Access*, *11*, 33336–33353.

Baron, V. (2020). *Méthodes d'identification de sources acoustiques paramétriques par mesures d'antennerie* (Unpublished doctoral dissertation). Université de Lyon.

Baron, V., Bouley, S., Muschinowski, M., Mars, J., & Nicolas, B. (2019). Drone localization and identification using an acoustic array and supervised learning. In *Artificial intelligence and machine learning in defense applications* (Vol. 11169, pp. 129–137).

Baron, V., Bouley, S., Muschinowski, M., Mars, J. I., & Nicolas, B. (2020). Acoustic localization and identification of drones with a disturbance source. In *Forum acusticum 2020* (pp. 3149–3154).

Bavu, É., Pujol, H., Garcia, A., Langrenne, C., Hengy, S., Rassy, O., ... Matwyschuk, A. (2022). Deeplomatics: A deep-learning based multimodal approach for aerial drone detection and localization. In *Quiet drones second international e-symposium on uav/uas noise.*

Bernardini, A., Mangiatordi, F., Pallotti, E., & Capodiferro, L. (2017). Drone detection by acoustic signature identification. *electronic imaging*, *29*, 60–64.

Blanchard, T., Thomas, J.-H., & Raoof, K. (2019). Acoustic signature analysis for localization estimation of unmanned aerial vehicles using few number of microphones. In *Matec web of conferences* (Vol. 283, p. 04002).

Blanchard, T., Thomas, J.-H., & Raoof, K. (2020). Acoustic localization and tracking of a multi-rotor unmanned aerial vehicle using an array with few microphones. *The Journal of the Acoustical Society of America*, *148*(3), 1456–1467.

Blass, M., & Graf, F. (2020). A real-time system for joint acoustic detection and localization of uavs. In *Quiet drones: International e-symposium on uav/uas noise (19-21 october 2020.*

Chiariotti, P., Martarelli, M., & Castellini, P. (2019). Acoustic beamforming for noise source localization–reviews, methodology and applications. *Mechanical Systems and Signal Processing*, *120*, 422–448.

Childers, D. G., Skinner, D. P., & Kemerait, R. C. (1977). The cepstrum: A guide to processing. *Proceedings of the IEEE*, *65*(10), 1428–1443.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Cousson, R., Leclere, Q., Pallas, M.-A., & Berengier, M. (2019). A time domain clean approach for the identification of acoustic moving sources. *Journal of Sound and Vibration*, *443*, 47–62.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... others (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Gong, Y., Chung, Y.-A., & Glass, J. (2021). Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*.

Herold, G., Kujawski, A., Strümpfel, C., Huschbeck, S., de Haag, M. U., & Sarradj, E. (2020). Detection and separate tracking of swarm quadcopter drones using microphone array measurements. In *Proceedings of the berlin beamforming conference (bebec), berlin, germany* (pp. 2–3).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Knapp, C., & Carter, G. (1976). The generalized correlation method for estimation of time delay. *IEEE transactions on acoustics, speech, and signal processing*, *24*(4), 320–327.

Koutini, K., Schlüter, J., Eghbal-Zadeh, H., & Widmer, G. (2021). Efficient training of audio transformers with patchout. *arXiv preprint arXiv:2110.05069*.

Lamotte, L., Baron, V., & Bouley, S. (2020). Uav detection from acoustic signature: Requirements and state of the art. In *Quiet drones: International e-symposium on uav/uas noise (19-21 october 2020).*

Leiba, R., Leclere, Q., & Julliard, E. (2022). Application of the cleant methodology to flyover noise measurements. In *9th berlin beamforming conference bebec 2022.*

Pujol, H., Bavu, E., & Garcia, A. (2021). Beamlearning: An end-to-end deep learning approach for the angular localization of sound sources using raw multichannel acoustic pressure data. *The Journal of the Acoustical Society of America*, *149*(6), 4248–4263.

Ramamonjy, A., Bavu, E., Garcia, A., & Hengy, S. (2018). Source localization and identification with a compact array of digital mems microphones. In *25th international congress on sound and vibration (icsv25).*

Schmidt, R. (1986). Multiple emitter location and signal parameter estimation. *IEEE transactions on antennas and propagation*, *34*(3), 276–280.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, *45*(11), 2673–2681.

Shi, L., Ahmad, I., He, Y., & Chang, K. (2018). Hidden markov model based drone sound recognition using mfcc technique in practical noisy environments. *Journal of Communications and Networks*, *20*(5), 509–518.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sun, Y., Li, J., Wang, L., Xv, J., & Liu, Y. (2023). Deep learning-based drone acoustic event detection system for microphone arrays. *Multimedia Tools and Applications*, 1–23.

Svanström, F., Alonso-Fernandez, F., & Englund, C. (2021). A dataset for multi-sensor drone detection. *Data in Brief*, *39*, 107521.

Utebayeva, D., Ilipbayeva, L., & Matson, E. T. (2022). Practical study of recurrent neural networks for efficient real-time drone sound detection: A review. *Drones*, *7*(1), 26.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Welch, P. (1967). The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics*, *15*(2), 70–73.

Wu, S., Zheng, Y., Ye, K., Cao, H., Zhang, X., & Sun, H. (2024). Sound source localization for unmanned aerial vehicles in low signal-to-noise ratio environments. *Remote Sensing*, *16*(11), 1847.