UNIVERSITÀ
DEGLI STUDI
DI UDINE

Università degli studi di Udine

Deep Acoustic Learning on Unmanned Aerial Vehicles for Real-Time Human and Drone Detection

Original

Availability:
This version is available http://hdl.handle.net/11390/1316185.3 since 2025-11-03T08:59:02Z

Publisher:

Published
DOI:

Terms of use:
The institutional repository of the University of Udine (http://air.uniud.it) is provided by ARIC services. The aim is to enable open access to all the world.

Publisher copyright

(Article begins on next page)

02 May 2026

Integrated Computer-Aided Engineering
XX(X):1–16
©The Author(s) 2025
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Deep Acoustic Learning on Unmanned Aerial Vehicles for Real-Time Human and Drone Detection

Andrea Toma* | **Daniele Salvati** | **Axel De Nardin** | **Silvia Zottin** | **Ivan Scagnetto** | **Giovanni Ferrin** | **Carlo Drioli** | **Gian Luca Foresti**

Abstract

In autonomous systems and robotics, acoustic signals offer valuable information for tasks such as acoustic source localization and recognition (LR), particularly in environments where visual sensing is limited. This paper investigates two unmanned aerial vehicles (UAVs)-based real-world scenarios that leverage acoustic scene awareness: (1) localization and recognition of human speech for search-and-rescue missions, and (2) detection and classification of other UAVs for counter-drone applications. To address these tasks, we design two deep learning models based on convolutional neural networks (CNNs) and a feature-based approach. These models process acoustic signals captured by two types of microphone arrays mounted on UAVs: a 4-microphone linear array and a 19-microphone spherical array. Each model performs direction-of-arrival (DOA) estimation and source classification under challenging ego-noise conditions using real-world datasets recorded in controlled experimental setups. We evaluate the models across different signal-to-ego-noise ratios and training configurations. Results show robust performance in both localization and recognition tasks, with about 6 degrees mean error and 7 degrees RMSE for DOA estimations in human speech scenario with multi-speaker classification accuracy till 0.95, and 3-5 degrees mean error and 7-11 degrees RMSE for DOA estimations in UAV sound scenario with multi-UAV classification accuracy till 0.98. This demonstrates the potential of deep acoustic learning for UAV-based scene understanding in complex operational environments.

Keywords

Features-based deep learning, voice localization and recognition, UAV localization and recognition, covariance matrix feature, CNN-based DOA estimation and acoustic signal recognition

1 Introduction

Sensing technologies are central to many civil applications, including event monitoring, object classification, search and rescue, and surveillance of both urban environments and disaster zones. The choice of sensors varies based on application needs and environmental constraints. In hostile environments, acoustic sensing and processing results to be suitable where other sensor data is inadequate, inapplicable, or unreliable. Examples include applications such as reconnaissance and surveillance with Unmanned Aerial Vehicles (UAVs) for contrasting intrusions (1-3), and search and rescue with adverse weather or even in hostile environments (4). In such conditions, speech and audio data with video fusion, like in (5-7), could be infeasible, even if multiple cameras are used as in (8), particularly in environments with occlusions or in disaster areas (9, 10) or in the cases the objective target is not in the region normally named field of view (FoV). In fact, in scenarios where the visibility is low, camera related sensors cannot help the data processing algorithms to provide significant performance. For this reason, applications based only on visual data do not guarantee the desired performance. In this sense, it is also not a proper choice to integrate visual and acoustic data because it increases the complexity of the algorithm without taking advantage of the use of cameras. In these conditions, the use of only acoustic data overcomes these limitations.

Recently, in the fields of deep learning (DL) and deep neural networks (DNNs) with studies like in (11-15), acoustic source localization and recognition (LR) algorithms have received increasing interest in the literature as it can be seen in (16-18). The usage of DL and DNNs in acoustic data processing could support acoustic scene awareness. In particular, regarding the application of UAVs, recent work analyzed the combined sound signals emitted by both UAV components, like propellers and electrical engines, and mechanical vibrations (18). That study was able to demonstrate that the combined noise has an acoustic signature that can be considered as sufficiently unique to recognize specific UAVs in a set of different UAV categories. DL and DNNs are also being investigated in different applications based on multi-channel acoustic processing like in (19-21) where the multichannel spectral phase information is used as input of a convolutional neural network (CNN) for the direction of arrival (DOA) estimation. The use of DL

Department of Mathematics, Computer Science and Physics (DMIF), University of Udine, Italy
{andrea.toma, daniele.salvati, axel.denardin}@uniud.it, zottin.silvia@spes.uniud.it, {ivan.scagnetto,giovanni.ferrin, carlo.drioli, gianluca.foresti}@uniud.it

***Corresponding author:**
Andrea Toma, University of Udine, Department of Mathematics, Computer Science and Physics (DMIF), Via delle Scienze 206, Udine, Italy. Email: andrea.toma@uniud.it

in audio processing applications for the improvement or the new design of multi-channel processing LR algorithms has been investigated only recently (19, 20, 22) and their use is being explored in many acoustic data-based applications like in (21, 23). Further investigations about localization of speakers based on DL networks can be found in (24-26).

In this study, the performance of CNN-based models is introduced for features-based LR tasks involving acoustic signals. The input stages process the phase values computed from the covariance matrix and the module of the frequency spectrum of the multichannel acoustic signals. The output stages perform regression where sound source localization (SSL), in terms of DOA estimation, is required and classification where acoustic source recognition, in terms of class estimation, is required. This is investigated in two real-world scenarios addressing the UAV-based sensing and processing of 1) voice emitted by humans and 2) propeller noise emitted by UAVs. In particular, in the experimental framework UAVs gather information from the environmental acoustic scene by using their own acoustic sensors consisting of microphone arrays with different configurations. In both Scenario 1 and Scenario 2, the task of the acoustic method is the LR, namely localization and recognition, of an acoustic source. Localization provides the DOA estimation of the propeller noise in Scenario 1 and of the human voice in Scenario 2. Recognition differentiates the type of human voice in Scenario 1 (normal or loud), and the sound of different UAVs in Scenario 2. This research is conducted in a context of autonomous systems where sensory data can be at the basis of acoustic scene awareness and decision-making. The sensory data are represented by acoustic signals. Two different real datasets with acoustic signals are collected for the two specific scenarios where different acoustic sources are considered as targets. The first dataset consists of voice signals recorded by means of a small 4-microphone linear array assembled on a UAV (Parrot Bebop quadcopter). The human is the target positioned in front of the UAV at different distances and angles and emits vocal signals in both normal and loud voice categories. To generate the second dataset, propeller noise recordings are produced by using two different UAVs (a DJI Matrice 200 and a F450 custom build quadcopter) as target. The data is collected by a medium-sized 19-microphone spherical array and includes two sound categories according to the target UAVs. In each of these two cases, the DOA and category of the acoustic sources are unknown parameters to be estimated. To benchmark the performance of the LR method, DOA estimation errors, sound recognition accuracy, and computing time are measured under different parameter settings.

In these scenarios, when the acoustic data is collected by microphone arrays with small inter-microphone distances installed on multirotor UAVs as in (27-29), the processing results to be challenging due to inadequate spatial resolution, signal-to-noise enhancement, and spatial information. The ego-noise that is generated by the propellers, electrical engines, and mechanical vibrations of the acquisition UAV that transports the microphone array also challenges the performance of the LR algorithm. Considering these limitations, an automatic acoustic source LR algorithm is implemented by using a data-driven approach. For this purpose, two alternative DL schemes based on CNNs are exploited for estimation of both DOA and category of the acoustic source. A feature-based approach for learning the models is considered.

This research motivation mainly consists in the analysis of the only acoustic scene through audio processing in such scenarios that still needs to be adequately investigated in the literature, especially in the framework of DL. In addition, applications in scenarios where video information is not reliable or available, the integration of audio processing with video analysis results to be unfeasible, and the algorithms merely based on audio processing should guarantee fully accurate results. In summary, this study introduces two real-world application scenarios based only on acoustic data processing, 1) a *speaking subject LR from UAVs* and 2) a *UAV LR from UAVs* both reproduced entirely at our university. Real acoustic sensors, UAVs, and acoustic sources are deployed for generating two real datasets representing both the acoustic scenes. Deep models in a data-driven approach are then investigated for the acoustic source LR. These models are capable of analyzing the multi-channel acoustic signals even in the presence of ego-noise. The evaluation of the LR models is conducted on both Scenario 1 and Scenario 2 datasets collected in real environments and the computation complexity has been investigated and compared with the conventional SRP-PHAT to evaluate the real-time ability of the method. The increase in the dataset diversity is also investigated for both Scenario 1, in contexts with 2 to 6 speakers, and in Scenario 2 with 6 different UAVs.

The rest of the paper is organized as follows: discussion on related work from the literature is provided in Sect. 2; two real-world scenarios based on LR from UAVs are introduced in Sect. 3; the general architecture of the proposed system with the logical structure of the algorithm is described in Sect. 4; acoustic sensors and experimental datasets based on acoustic data related to the two scenarios are presented in Sect. 5; experiments and results of the LR algorithm based on the two DNN models for a data-driven approach are shown in Sect. 6; a detailed analysis of the outcomes is then reported in Sect. 7; conclusion of this study and future work are the final parts of the manuscript in Sect. 8.

2 Related work

From the analysis of the state-of-the-art, four categories, among others, have been identified according to the acoustic application the research was conducted on: 1) human voice localization, 2) human voice recognition, 3) UAV localization, and 4) UAV recognition.

- *The first category* concerns voice recognition from UAVs. A voice recognition-based detection system for search and rescue in large-scale disasters like huge earthquakes is presented in (30). A speaker installed on the UAV makes sounds for victims to react, and victims are detected by the recognizer that captures the reaction voice of the victims. The captured acoustic signal consisting of the human voice also contains ego-noise and other outdoor environmental sounds.

A speaker identity verification through voice recognition is presented in (31) for security and privacy of voice-controlled UAVs. Features are first computed from the audio

signals as Mel-Frequency Cepstral Coefficients (MFCC) and then a feature matching is applied. The authors employed DL as a soft computing tool capable of enabling intelligent systems that mimic human behavior.

A sound source separation and identification algorithm for processing noise-contaminated acoustic signals is addressed in (32). Audio data is gathered with a microphone array embedded in a UAV to detect people's voice quickly and widely in a disaster situation. The authors propose a Partially-Shared DNN (PS-DNN) which can learn with a small amount of annotated data.

- *The second category* is the voice localization from UAVs. An algorithm proposed for drone-based search and rescue operations during disaster management is investigated in (33) to estimate the coordinates of the speech source. Specifically, the person who needs help (acoustic source) screams, and the drone with an onboard microphone array gathers the vocal signal originated by the human. An approach is proposed to analyze the captured audio signals for acoustic source localization. The ego-noise, consisting of the noise related to the motion of the UAV, the noise produced by the UAV propellers and by the electrical engines of the UAV, and other stationary structural noise, challenges the performance of the SSL algorithm based on the time difference of arrival.

The publicly-available DREGON dataset introduced in (34) was developed for research purposes in SSL. It was generated using a microphone array embedded in a UAV. The dataset contains both clean and noisy in-flight audio recordings annotated with the 3D position of the target sound source. It can be used for emerging tasks of UAV-embedded SSL. The study conducted on the dataset showed promising performance of the localization algorithm of a broad-band acoustic source in presence of high level of noise, whilst speech localization remained challenging under extreme noise levels.

The authors of the work related to the IEEE Signal Processing Cup 2019 Student Competition in (35), highlight that, although UAVs that are equipped with an acoustic sensor like a microphone array could largely help localization of people in cases of emergency where video acquisition is severely limited due to a reduced or absent visual information caused by limited lighting conditions, e.g. nocturnal or in fog acquisitions, or by the presence of obstacles that limits the FoV, UAV-based acoustic localization has not been investigated sufficiently, yet.

- *The third category* is related to UAV recognition. The drone recognition system studied in (36), based on audio signals generated by UAVs, exploits MFCC as the audio feature and uses both a Support Vector Machine model (SVM) and a CNN for recognizing the audio generated by UAVs. A small UAV audio dataset was created.

UAV identification in extreme environmental conditions and large dataset requirements is addressed in (37) where an ensemble DL framework is proposed to contrast unauthorized or malicious UAVs. The UAV classification is based on hybrid synthetic and deep features computed from acoustic signals fused with data from other sensors.

UAV recognition in (38) is performed first using CNN models trained by audio spectrograms in addition to other data and then the CNN output probability is processed by multinomial logistic regression. For the experiments, the datasets are both collected by field measurements of real UAVs using audio microphone for the acoustic processing and obtained from open online repositories.

- *The fourth category* corresponds to UAVs localization. The Drone Acoustic Detection System (DADS) from the Stevens Institute of Technology uses the DOA and localization to track UAVs based on their propeller noise, (39). The system is based on microphones arranged in a tetrahedron. A 16-channel two-tier cross array, the OptiNav 40-microphone phased array, and parabolic and shot gun microphones were also considered. While the Multirotor UAVs employed in testing were DJI models Phantom 4, M600 and S1000.

A UAV acoustic source localization algorithm based on ESPRIT with Toeplitz matrix reconstruction is proposed in (40). Indoor and outdoor tests were conducted with a 12-channel spherical microphone array and a circular MEMS microphone array designed by the authors.

DOA estimation of an intruding UAV through its acoustic signature from harmonics extracted from the received sound signals is the objective in (41). The proposed method first estimates the harmonic frequencies corresponding to the frequency domain acoustic signal of the UAV are first estimated. A classifier for multiple signals is then applied to compute the estimates of the DOAs of the set of harmonics. A weighted sum of the estimated DOAs is then computed as the drone's DOA estimate where the weights are proportional to the energy of the harmonics.

Recent and detailed overviews on SSL in drone audition are also provided in (42, 43).

As mentioned above, this study introduces two real-world application scenarios based only on acoustic data processing, 1) a *speaking subject LR from UAVs* and 2) a *UAV LR from UAVs*, which are described later in the manuscript. These scenarios were reproduced at our university to generate two real datasets, according to them. Real acoustic sensors, UAVs, and acoustic sources are deployed. This allowed the proposed DL models thought for acoustic source LR to be experimented. In fact, two deep models that are capable of extracting information from the acoustic scenes on a data-driven basis are proposed in this research.

The motivation for considering the analysis of the only acoustic scene is that application of multi-channel audio processing for LR tasks needs more investigation in the literature especially in the framework of the rapid developing DL. Moreover, this research addresses the LR problem in scenarios where video information is not reliable or available, like when based on nocturnal or in fog acquisitions. In these cases, the only audio-based algorithms should provide fully accurate analysis of the surrounding scene. An iterative diagonal unloading (IDU) beamforming based on the identification of the dominant signal for DOA estimation in acoustic multi-channel signal processing is studied in (44).

Each cited work addresses peculiar scenarios, models, and issues, we conclude this section with a brief account of recent works which are more similar to our approach. Hence, the comparison is restricted to recent CNN-based solutions previously mentioned. Among those contributions, a set of studies propose systems for predicting the DOA of sound

sources. For instance, the approach in (19) focuses on a search and rescue application that employs a simple stereo microphone onboard a UAV, and proposing a parametric multi-channel Wiener filter to cope with the UAV ego noise. Then, power level-based features are extracted (e.g., PLR, PLD, PLS) and fed into a CNN to predict the DOA of the sound source. The CNN architecture is a chaining of 6 convolutions blocks with two-dimensional convolutions, batch normalization, ReLU activation, and max-pooling layer and two fully connected blocks. In (20), in the context of multi-speaker DOA estimation, the issue of lack of large amounts of labeled data for training is addressed, by leveraging on data augmentation and weakly-supervised domain adaptation. Simulation is used to generate source domain data, while collected real data are annotated with the number of sound sources as weak labels. The latter are then augmented by mixing single-source segments. Finally, weakly-supervised domain adaptation is applied to models pretrained on the simulated data. Such approach, according to the experiments with real robot audio data carried out by the authors, allows one to achieve similar performance to the fully-labeled real data scenarios. The problem of determining DOA of sound sources as a classification problem is addressed also in (24) where two CNNs are used to infer elevation and azimuth of sound sources, leveraging on Spherical Harmonic Decomposition. Indeed, the latter allows for the extraction of two sets of features containing information about elevation and azimuth of the sound source. Finally, the authors of (26) focus on DOA estimation in noisy and reverberant environments, by exploiting DNNs to identify speech dominant time-frequency units with a relatively clean phase. This is particularly useful for DOA estimation with a DNN trained using only monaural spectral information. Hence, this yields a model directly applicable to microphone arrays with diverse geometries.

From the above mentioned studies, it emerges that one the novel contributions of this work is providing an approach leading to solutions that combine sound classification and DOA estimation in a complete LR system (with two kinds of not trivial microphone arrays). Moreover, the associated CNNs are rather lightweight, using 3-4 convolutions for the first Scenario (two separate CNNS for sound classification and DOA estimation) and 7 convolutions for the second Scenario (exploiting a data fusion strategy to combine both kinds of predictions). This is a consequence of the effectivity of the chosen features.

To conclude, our study mainly focuses on the application of the LR algorithm to real scenarios featuring multi-channel acoustic data for two different targets, human voice and UAV sound. We present two datasets collected according to the scenarios where two microphone array configurations (4-microphone linear with 16kHz sample-rate and 19-microphone spherical with 48kHz) have been employed. In the revised version of our study, multiple UAVs and voices have been employed for an increased data diversity. The algorithm is made in that way to be suitable for embedded applications. Complexity analysis and comparison with the conventional SRP-PHAT are conducted to verify the performance of the DNN networks.

3 Two real-world scenarios with UAVs based on acoustic signal processing

In this section, the two real-world scenarios for analyzing the acoustic scene, 1) a *speaking subject LR from UAVs* and 2) a *UAV LR from UAVs*, are described in details. The term "*real-world*" is basically intended as not simulated scenarios and refers to applications that can be found in the reality (not hypothetic scenarios). The two datasets and the extended versions are collected in real environments.

3.1 Scenario 1: acoustic data-based speaking subject LR from UAVs

In Scenario 1, a UAV equipped with an onboard microphone array collects acoustic data from its environment. The target source is a human emitting vocal signals, which are captured alongside ego-noise generated by the UAV itself. This configuration enables the system to detect and localize human presence through voice, supporting applications such as human-UAV interaction and search-and-rescue missions.

The acoustic signal received at the microphone array consists of two main components: the voice sound emitted by the target source (human) and the propeller noise emitted by the acquisition UAV where the microphone array is mounted (ego-noise). The collected signals can then be processed through signal processing methods for different tasks, such as detection, recognition, and localization of the target human (sound source) by means of the intrinsic characteristics of the voice. Possible application fields can be the human-UAV interaction and search and rescue where the acquisition and processing of acoustic sensory information for LR of human voice may enable situation awareness.

3.2 Scenario 2: acoustic data-based UAV LR from UAVs

An acquisition UAV gathers information from the surrounding acoustic scene by means of the microphone array the UAV is equipped with. The acoustic sensor is mounted onboard the UAV. It is able to gather the acoustic signals that come from sources in the surrounding area. In this scenario, the acoustic source is a UAV that performs flight maneuvers in the observed area. It generates propeller noise that propagates in the surrounding area.

The acoustic signal received at the microphone array consists of two main components: the propeller noise emitted by the target source (UAV) and the propeller noise emitted by the acquisition UAV where the microphone array is mounted (ego-noise). The collected signals can then be processed through signal processing methods for different tasks, such as detection, recognition, and localization of the target UAV (sound source) by means of the intrinsic characteristics of its propeller noise. Possible application fields can be the UAV-UAV interaction and counter-UAV, Fig. 1, where acquisition and processing of acoustic sensory information for LR of UAVs may enable situation awareness.

Acoustic data-based speaking subject LR from UAVs (scenario 1) and acoustic data-based UAV LR from UAVs (scenario 2) are both the real-world scenarios of the research conducted in this work. In both of them, the major cause of performance degradation is the ego-noise generated by

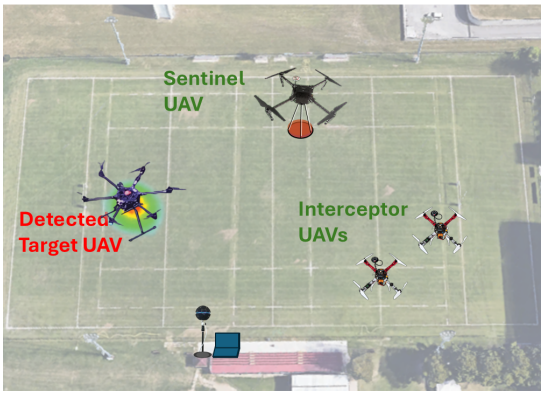

Figure 1. An example of Scenario 2 thought for counter UAV applications where a sentinel UAV and interceptors UAVs contrast an intruder UAV by acquiring information from the surrounding acoustic scene through microphone arrays

the propellers. Several studies on UAV propeller noise characterization can be found in the literature, such as in (45-48).

4 General architecture of the proposed system

The general system architecture is represented in Fig. 2 where the physical representation of the system consisting of an acoustic source such as a UAV or a human and a UAV equipped with a multi-microphone array (upper part of Fig. 2) is shown along with the logical structure of the algorithm for the DOA estimation $\widehat{\Theta}$ (localization) and class prediction $\widehat{C}$ (recognition) tasks computed through a CNN-based deep network (bottom part of Fig. 2). An alternative scheme could have two different branches without the fusion layer to keep separated the two paths: DOA estimation based on the feature computed as the phase of each element of the covariance matrix $\measuredangle\widehat{\mathbf{\Phi}}(k,f)$, where k denotes the time index and f denotes the frequency bin, and the class prediction based on the feature computed as the module of the signal spectrum $|X(f)|$, where f denotes the frequency bin. Both variants are investigated in this research. A formal and detailed description of the different parts of the logical structure of the algorithm is provided in the following sections from the acoustic array formulation to the CNN-based DNN models.

4.1 Acoustic array formulation

The acoustic array formulation considers a microphone array with M omnidirectional sensors and an acoustic source in the far-field. $\mathbf{\Omega} = [\Theta]$ is the DOA of the acoustic wave in the array, where Θ denotes the azimuth angle. In a single-source scenario, the data model in the short-time Fourier transform (STFT) domain, relative to the multi-channel signal, can be expressed as:

$$\mathbf{x}(k,f) = \mathbf{a}(f,\mathbf{\Omega})S(k,f) + \mathbf{v}^e(k,f) + \mathbf{v}(k,f) \quad (1)$$

where k is the frame time index, f denotes the frequency bin, $\mathbf{a}(f,\mathbf{\Omega})$ is the array steering vector for the source direction $\mathbf{\Omega}$, $S(k,f)$ represents the source signal relative to the reference sensor, $\mathbf{v}^e(k,f)$ is the non-stationary UAV

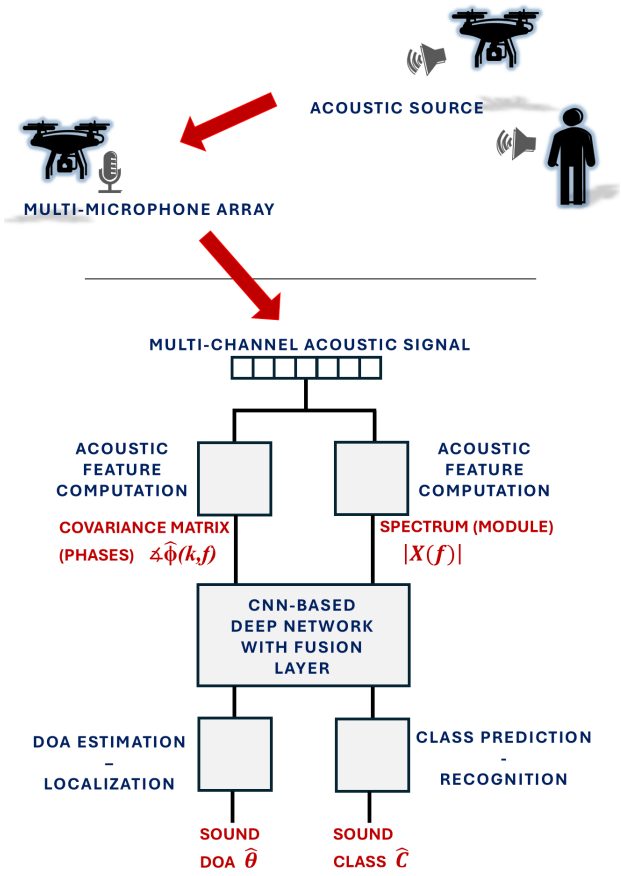

Figure 2. General architecture of the system. Physical representation of the system (top) and logical structure of the algorithm (bottom)

ego-noise consisting of multiple narrowband harmonic noise generated by both motors and broadband aerodynamic noise of the UAV propellers, and the term $\mathbf{v}(k,f)$ represents the additive noise as spatially white Gaussian noise characterized by mean and variance values equal respectively to zero and σ^2 for all the mirophonic sensors.

The covariance matrix of the array signal can be computed as $\mathbf{\Phi}(k,f) = E\{\mathbf{x}(k,f)\mathbf{x}^H(k,f)\}$, which is symmetric and positive definite, where $E\{\cdot\}$ denotes the mathematical expectation. The covariance matrix is computed in the frequency range (f_{min}, f_{max}). The elements in the matrix $\mathbf{\Phi}(k,f)$ denote the correlation between pairs of microphones. Specifically, in array processing, a covariance matrix represents the correlations between signals received by different elements of an array, such as microphones, and is essential for characterizing desired signals and suppressing undesired interference and noise. The covariance matrix is derived from the output vector of the array which is often complex-valued representing the signals received by the array. It is organized in a square matrix that contains the variances of each sensor output along its main diagonal and the covariances between pairs of sensors outputs off the diagonal. The covariance matrix can be used in algorithms like interference and noise characterization, adaptive filtering, target localization, and data-driven estimation. Since the covariance matrix is often unknown in practical applications, it has to be estimated. Its

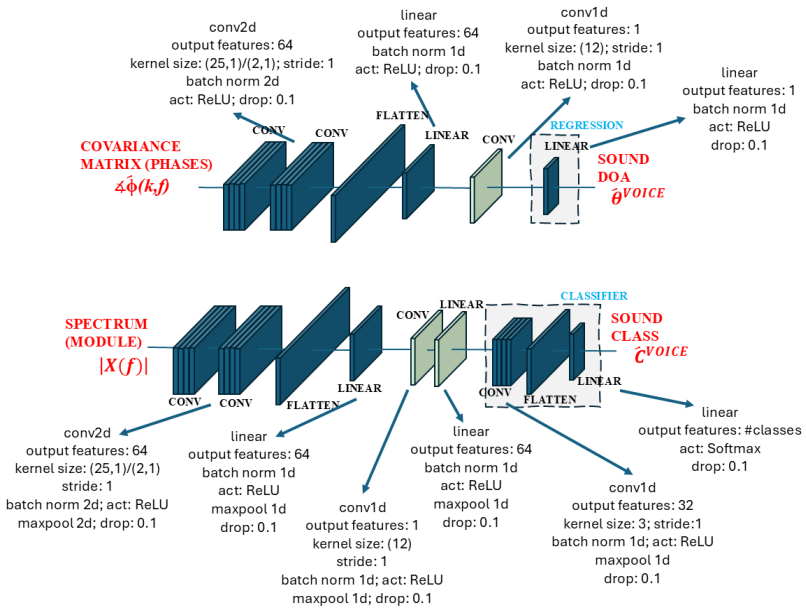

Figure 3. Features-based DNN scheme for data-driven acoustic source LR designed for the Scenario 1. The layer composition and parameters are visualized in the scheme

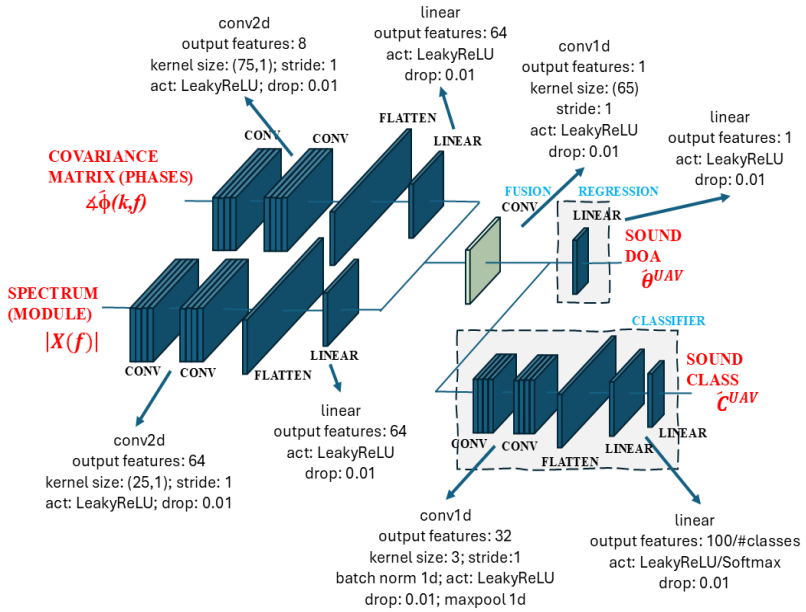

Figure 4. Features-based DNN scheme for data-driven acoustic source LR designed for the Scenario 2. The layer composition and parameters are visualized in the scheme

estimated model can be computed by averaging over blocks of the multi-channel signal:

$$\widehat{\mathbf{\Phi}}(k,f) = \frac{1}{B}\sum_{k_b=0}^{B-1} \mathbf{x}(k-k_b,f)\mathbf{x}^H(k-k_b,f) \qquad (2)$$

where B denotes the mean number of snapshots and f is in the range (f_{min}, f_{max}). However, due to the finite sample size (number of snapshots), to the signal model mismatches, and to the non-stationary nature of the source, there is always a certain mismatch between the estimated and the true covariance matrix. Further details can be found in (49-51).

On the other hand, the spectrum $\mathbf{X}(f)$ of the multi-channel signal can be computed for each snapshot.

4.2 Acoustic features used in the DNN models

The estimated covariance matrix $\widehat{\mathbf{\Phi}}(k,f)$ consists of $M \cdot M$ complex values, where M is the number of microphones in the array. Phase values of each element in the matrix $\widehat{\mathbf{\Phi}}(k,f)$ can be computed and the resulting matrix can be denoted by $\measuredangle\widehat{\mathbf{\Phi}}(k,f)$ with $\measuredangle$ denoting the element-wise phase operator. The phase values provide acoustic source delay information between each couple of microphones.

The spectrum $\mathbf{X}(f)$ of the discrete-time multi-channel acoustic signal $\mathbf{x}(k)$ is computed as $FFT(\mathbf{x}(k))$, where FFT stands for Fast Fourier Transform. Its module $|X(f)|$ provides the frequency content of the audio signal and can be

used in the sound classification task where only 1 channel of the acoustic signal is considered due to the invariance of the spectrum over the channels.

4.3 Acoustic tasks performed by the DNN models

The localization part of the algorithm provides estimations, denoted by $\widehat{\Theta}^{VOICE}$ in Scenario 1 and $\widehat{\Theta}^{UAV}$ in Scenario 2 , of the DOA of the acoustic signals. SSL is a regression problem since the DOA values belong to the continuous range. In the proposed study, the DOA estimation refers to the azimuth angle, on the horizontal plane, between the acoustic source and the microphone array. In the Scenario 1, the localization method provides predictions $\widehat{\Theta}^{VOICE}$ of the DOA of the acoustic signal emitted by the human speech (human voice localization). In the Scenario 2, it provides predictions $\widehat{\Theta}^{UAV}$ of the DOA of the acoustic signal emitted by the target UAV (UAV localization).

The recognition part of the algorithm provides estimations, denoted by $\widehat{C}^{VOICE}$ in Scenario 1 and $\widehat{C}^{UAV}$ in Scenario 2, about the acoustic signal according to its acoustic characteristics. This is a classification problem. When two classes should be differentiated, a binary recognition method is employed. It produces the probability of recurrence for each of the two classes, namely it consists of both the correct decision probability and the false alarm probability. According to these two probability values, the method chooses the class with the highest probability of occurrence. When binary recognition is performed, the method decides either 0 or 1. In the Scenario 1, normal voice (N) and loud voice (L) should be differentiated, $\widehat{C}^{VOICE} \in \{\widehat{C}^{N}, \widehat{C}^{L}\}$, according to the acoustic characteristics or acoustic signature of the speech signal emitted by a human. In the Scenario 2, the two classes correspond to two different UAVs, Type 1 UAV ($T1$) and Type 2 UAV ($T2$), and they are differentiated, $\widehat{C}^{UAV} \in \{\widehat{C}^{T1}, \widehat{C}^{T2}\}$, according to the acoustic characteristics or acoustic signature of the noise emitted by the UAV propellers.

4.4 CNN-based DNN models

In this study, two deep network architectures are investigated and proposed. The first structure in Fig. 3 is employed in Scenario 1. The regression network is separated from the classification network. Sound DOA estimation is based on the covariance matrix feature. Sound classification uses the 1-channel spectrum module feature. Both branches pass through a convolution layer before the corresponding head networks for regression and classification, respectively. The layer composition and corresponding parameters are also listed in the scheme.

The second structure in Fig. 4 is employed in Scenario 2. Two parallel branches process the input features, namely the phase values computed from the estimated covariance matrix, $\angle\widehat{\mathbf{\Phi}}(k,f)$, and the module of the 1-channel acoustic spectrum, $|X(f)|$, through convolution, flatten and linear layers. A convolution fusion layer joins the two branches just before the two head networks for sound DOA estimation and sound classification, respectively. The sound DOA estimation is obtained by regression, and the sound classification is implemented by means of a classifier head.

(a) (b)

Figure 5. a) Sony PlayStation Eye (PS3 Eye) where the 4-microphone linear array is visible on top of the device and b) Zylia ZM-1 consisting of 19 microphones organized in a spherical array

The layer composition and corresponding parameters are also listed in the scheme.

Furthermore, the network and learning parameter values are also listed in detail in Sect. 6.3 for the ease of understanding of the experiments.

The motivation of introducing these two deep networks is to largely reduce the computational complexity of the algorithm by substituting part of the heavy mathematical computation relative to multi-channel array processing with a data-driven learning deep network model without any performance reduction of the LR algorithm.

4.5 DOA estimation and Signal recognition

The task of the algorithm is the DOA estimation to predict the azimuth angle, $\widehat{\Theta}^{VOICE}$ or $\widehat{\Theta}^{UAV}$, of the acoustic signal and the signal recognition to differentiate the categories, $\widehat{C}^{VOICE}$ or $\widehat{C}^{UAV}$, of the acoustic signal. This study follows a deep learning approach to perform the two tasks, as detailed through the scheme in Fig. 2 and the details on the computation are explained throughout the section from the acoustic array formulation to the acoustic features, acoustic tasks, and the DNN models. The first task is implemented as a regression problem, and the latter task is implemented as a classification task.

5 Acoustic sensors and experimental datasets

Experiments with acoustic signals have been conducted with the objective of validating and demonstrating the proposed method. In this section, two microphone arrays, a *4-microphone linear array* and a *19-microphone spherical array*, and two datasets of acoustic signals, a *human voice generated by speech dataset* and a *UAV noise generated by propellers dataset*, are introduced as they were used in the experiments. The three UAVs used in the experiments and mentioned in this section are: 1) a *Parrot Bebop* (52), 2) a *DJI Matrice 200* (53), and 3) a *F450 custom build quadcopter* (54).

5.1 Acoustic sensors

1) *4-microphone linear array*. The Sony PlayStation Eye (PS3 Eye) (55) is a small and lightweight device, as shown in Fig. 5(a). In addition to a camera, the PS3 Eye features a built-in four-capsule linear microphone array, with which technologies for multidirectional voice location tracking, echo cancellation, and background noise suppression can

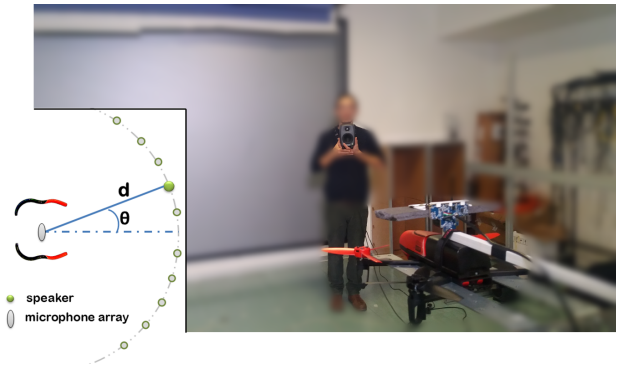

Figure 6. Setup of the Scenario 1 where the acoustic signal generated by the human voice is gathered by the 4-microphone array PS3 Eye and main scheme of the experiment (on the left-bottom part of the immage) with the positions of the target human (speaker). The horizontal distance values (d [meters]) are 2, 3, and 4 meters. The azimuth angles values (Θ [degrees]) are $[-5, 5]$, $[-10, 10]$, $[-15, 15]$, $[-20, 20]$, and $[-25, 25]$ degrees

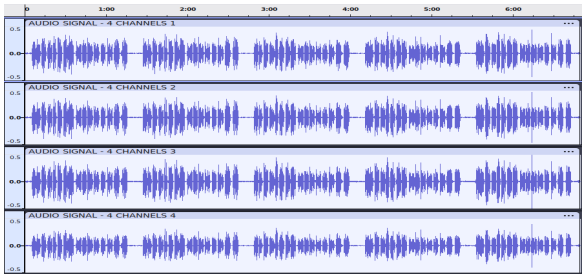

Figure 7. Acoustic signal generated by human voice of a target speaker and collected by the PS3 Eye sensor all the 4 channels are visualized in the image). The time duration of the visualized recording is 410 seconds and the sampling rate is 16 kHz

be employed. The acoustic array is designed to have the microphones equidistant with inter-microphone distance and total length equal to 2cm and 6cm, respectively. The peripheral can be used in applications such as speech recognition (SR) and audio chat in noisy environments without the use of a headset. The PS3 Eye microphone array operates with each channel processing 16-bit samples at a sampling rate of 16KHz, and a signal-to-noise ratio of 90dB. Several technologies are available for the PS3 Eye. Among these are the PSVR (PlayStation Voice Recognition), a SR library that is intended to support about 20 different languages.

2) *19-microphone spherical array*. Zylia ZM-1 (56) is a multichannel microphonic device used for acquisition. As a compact spherical array, it consists of 19 digital omnidirectional microphonic capsules (XENSIV) based on technology MEMS from German-based Infineon Technologies forming a sphere whose diameter is 9 cm, Fig. 5(b). The nominal signal-to-noise ratio is 69dB, the dynamic range is 105dB, and the output linearity is guaranteed up to 130dB. Zylia ZM-1 is capable of capturing the entire surrounding acoustic scene in 3D. The sample rate and the resolution are 48kHz and 24 bit, respectively. The acoustic gain is adjustable in the range from 0 to 70dB. The microphones do not require re-calibration thanks to their matched and constant over-time parameters. The front of the sphere is marked with a painted dark-red dot.

5.2 Acoustic datasets

The two experimental datasets described in this section and the related annotation files are freely available at the link (57).

1) *Data consisting of human voice generated by speech*. The dataset was gathered according to the Scenario 1 introduced in Sect. 3.1 and corresponds to the experimental setup represented in Fig. 6. It features a human subject whose pre-recorded vocal signals were alternatively emitted by two acoustic loudspeakers placed frontally with respect to the drone. The two loudspeakers were positioned symmetrically with respect to the frontal direction of the drone and change their positions, in terms of angle and distance, uttering the same sentence one at time with two different voice modes: normal and loud. A set of 30 positions was used for each of the two categories (normal voice and loud voice) according to the following angles (Θ[degrees]) and distances (d[meters]) in a 2D space: $[-5, 5]$, $[-10, 10]$, $[-15, 15]$, $[-20, 20]$, and $[-25, 25]$ degrees at 2, 3, and 4 meters. The voice sequence was repeated at each of these positions.

An example of acoustic signal generated by propeller noise of a target UAV (Parrot Bebop) and collected by the PS3 Eye is shown in Fig. 7 where all the 4 channels are visualized. The time duration of the visualized recording is 410 seconds and the sampling rate is 16 kHz.

The ego-noise is a pre-recorded signal previously obtained by recording the propeller noise of a Parrot Bebop UAV collected by the 4 microphone array PS3 Eye, shown Fig. 8. After SNR calibration, in the synthesis method the ego-noise is summed up by time domain superposition to the signals relative to the targets previously described. The signal-to-ego-noise ratio (SNR) can be set according to the specifications of the experiment. To increase the diversity of the dataset, a software tool for producing different voices from an acoustic recording in the Scenario 1 previously described has been utilized. It applies acoustic transformations to the voice signal to change its characteristics according to different voices resulting in a number of speakers (more than 10 speakers can be obtained). The drawbacks of this method are the loss of spatial information, normal/loud characteristics, and channel separation. The extended diversity dataset is employed for testing the multi-voice classification performance of the DNN-based recognition algorithm.

2) *Data consisting of UAV noise generated by propellers*. The dataset was collected according to the Scenario 2 introduced in Sect. 3.2 and corresponds to the experimental setup represented in Fig. 9. The acoustic signals, sequentially emitted by UAVs that generate propeller noise as acoustic sources hovering at 6 specific positions defined by their coordinates, are recorded by using the Zylia. Due to its weight, it is still unsafe for a lightweight UAV to transport Zylia. The altitude of the spherical acoustic array, measured in relation to the ground, is equal to 1.60m. Two UAVs with different characteristics, in the role of unidentified acoustic sources, fly in the area near the acoustic

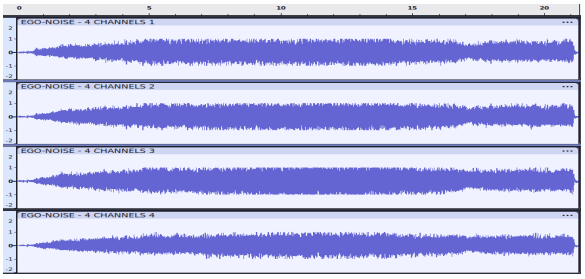

Figure 8. Ego-noise signal generated by propeller noise of a hovering Parrot Bebop UAV and collected by the PS3 Eye array mounted on the UAV (all the 4 channels are visualized in the image)

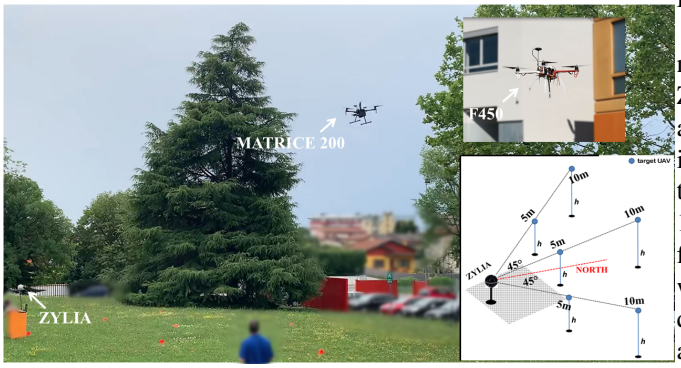

Figure 9. Setup of the Scenario 2 where the acoustic signal generated by the propellers of hovering UAVs is gathered by the 19-microphone array Zylia ZM-1 and main scheme of the experiment (on the right-bottom part of the immage) with the positions of the target UAVs. The 3D distance values (d [meters]) are 5.016 and 10.179 meters. The azimuth angles values (Θ [degrees]) are -45, 0, and $+45$ degrees. The elevation angles values (ϕ [degrees]) are 4.574 and 10.758 degrees

sensor. First, the DJI Matrice 200, and afterwards the F450, separately at two different acquisition sessions. The two UAVs were in hovering mode during the acquisition time at the 6 coordinate points for each of the two UAVs according to the following azimuth angles (Θ[degrees]) and distances (d[meters]) in a 3D space: -45, 0, and $+45$ degrees at 5.016 and 10.179 meters. The elevation angles (ϕ[degrees]) were: 4.574 and 10.758 degrees.

The front direction relative to the Zylia is the reference direction. The experimental setup that reproduces the real acoustic scene is shown in Fig. 9. The Zylia is placed in the middle of the meadow at the university campus close to the gym building. To increase the diversity of the dataset, four additional UAV targets have been employed in the same way as previously described according to the Scenario 2 specifications. The 4 additional UAVs were 1) an Aurelia X6 Hexacopter, 2) a DJI Phantom 4 Quadcopter, 3) a DJI Mini 4K Quadcopter, and 4) a Yuneec H520E Hexacopter resulting in a total of 6 UAV targets (considering the two UAVs described previously). This results in a multi-class problem and the extended diversity dataset is employed for testing the multi-UAV classification performance of the DNN-based recognition algorithm. Concerning the environmental noise, reverberation from obstacles, and other challenging conditions, the dataset acquisitions have been conducted in open spaces (close to roads, industrial buildings, trees, and so forth) with outdoor conditions where reverberation and a number of environmental sounds from the surrounding area were actually captured, too, and they surely impacted the LR algorithm performance. To create the dataset with acoustic signals consisting of propeller noise recordings according to the six configurations for the hovering UAVs, the audio signals collected by using the spherical acoustic array are recorded as mp4 file. The recording time for each of the configurations is 19-30 seconds, the audio gain of the Zylia is set to 0dB, the data format is float with 32 bits, the FFT length is equal to 2048 samples, and the sample-rate is 48kHz. The sounddevice Python module is used in the registration software.

An example of acoustic signal generated by propeller noise of a target UAV (Matrice 200) and collected by the Zylia ZM-1 is shown in Fig. 10 where 8 out of 19 channels are visualized. The time duration of the visualized recording is 33 seconds and the sampling rate is 48 kHz. Tests on the Scenario 2 setup were conducted with the Zylia ZM-1 mounted below a F450 UAV, as in Fig. 11(a), and used for ego-noise measurements. In addition, real ego-noise was also collected by flying a UAV (an Holybro X500 V2 quadcopter) as a different ego-noise source over the Zylia at a safety distance from the spherical array as in Fig. 11(b). In fact, since it is unsafe for the F450 to transport Zylia and both lighter spherical arrays and planar arrays have not comparable quality and performance with respect to the Zylia, the hovering state of the acquisition UAV is mimicked both by fixing the UAV to a stable support and by flying the UAV over the Zylia. The ego-noise is in this way a pre-recorded signal previously obtained by recording the propeller noise of the F450 UAV collected by the 19 microphone array Zylia. An example is shown in Fig. 12. After SNR calibration, in the synthesis method the ego-noise is summed up by time domain superposition to the signals relative to the targets previously described. This mimics the hovering state of the F450. The SNR can be set according to the specifications of the experiment.

6 Experiments and results

6.1 HW and SW processing platforms

A Laptop with i7-11800H CPU @ 2.30GHz, 16 GB RAM, NVIDIA GeForce RTX 3050 Ti Laptop GPU (4 GB), and 64-bit Windows 10 Pro OS was used for data collection and store. A desktop workstation with i9-10920X CPU @ 3.50GHz, 64 GB RAM, NVIDIA GeForce RTX 3090 GPU (24 GB), and 64-bit Windows 10 OS was used for processing and train/validation/test of the deep neural model.

Python 3.11.4 with its library Pytorch 2.0.1+cu117 was used for the experiments on the deep network consisting of training, validation, and testing of the network. The Python libraries Soundfile and Sounddevice were used for reading and extracting the acoustic data.

6.2 Datasets of acoustic signals

The datasets used in the experimental study cover both the Scenario 1 for *acoustic data-based speaking subject LR from*

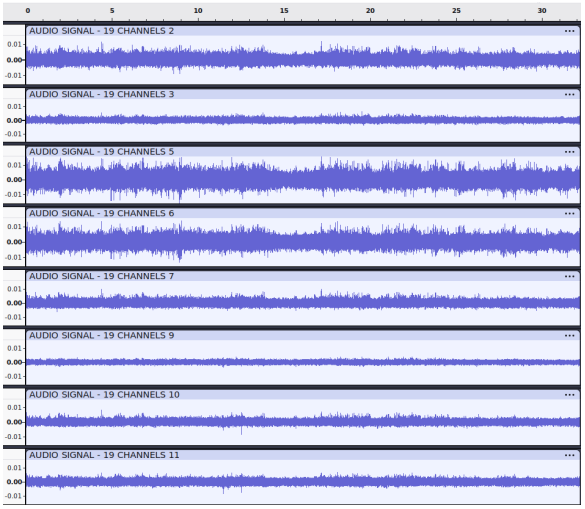

Figure 10. Acoustic signal generated by propeller noise of a target UAV (Matrice 200) and collected by the Zylia ZM-1 sensor (only 8 out of 19 channels are visualized in the image). The time duration of the visualized recording is 33 seconds and the sampling rate is 48 kHz

(a) (b)

Figure 11. Zylia ZM-1 a) mounted below a fixed F450 quadcopter b) below a flying Holybro X500 V2 quadcopter for testing the Scenario 2 setup and ego-noise measurements

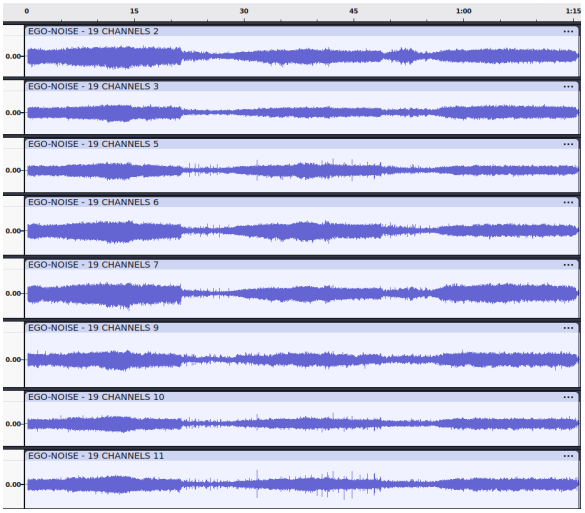

Figure 12. Ego-noise signal generated by propeller noise of a hovering F450 UAV and collected by the Zylia ZM-1 array mounted on the UAV (only 8 out of 19 channels are visualized in the image)

UAVs (Sect. 5.2-1) and the Scenario 2 for *acoustic data-based UAV LR from UAVs* (Sect. 5.2-2).

6.3 Signal and learning parameters

Values of the signal and learning parameters used for the experiments are listed in Table 1.

Table 1. Signal and Learning parameters

	Scenario 1	**Scenario 2**
Signal parameters		
Number of channels	4	19
Sample rate	16000	48000
Frame length	2048	
Windowing	Hanning	
f_{min}/f_{max} (Hz)	20/16000	
SNR	2:1 and 1:1	
Learning parameters		
Number of epochs	520	390
Batch size	32 and 64	32 and 64
Optimizer (regression)	Adam	Adadelta
Optimizer (classification)	Adam	SGD
Learning rate (regression)	0.001	0.001
Learning rate (classification)	0.001	0.01
Loss function (regression)	SmoothL1Loss	
Loss function (classification)	CrossEntropyLoss	
Train-Val-Test data splits	0.5-0.2-0.3	

6.4 Experimental results

The outcomes of the experiments for the two scenarios are introduced in this section.

Scenario 1: (speaking subject LR with 4-channel signals) the system was tested under two signal-to-ego-noise ratios (SNR1 = 2:1 and SNR2 = 1:1) and two batch sizes (32 and 64). Results demonstrated consistent localization accuracy across settings. The mean azimuth error varied between 5.78° and 6.26°, with median errors in the range of 5.45° to 6.16°. The RMSE values remained under 7.4°, indicating reliable DOA estimation even under high ego-noise. Voice classification accuracy ranged from 81% to 89%, showing minor sensitivity to batch size or SNR. Overall, the model retained strong performance despite noise interference.

The conventional SRP-PHAT method for acoustic target localization was applied to the human voice-based dataset in Scenario 1 to compare the performance with our DNN-based method. The time domain variant of the SRP-PHAT was considered with 100 and 200 grid points. The results are organized in Table 3 where two two ego-noise levels are considered.

The classifier performance when applied to the increased diversity dataset with 2 to 6 different voices can be seen from the test accuracy curve in Fig. 13 obtained by setting the batch size to 32 and the ego-noise level equal to the signal level. For this experiment, the MFCC coefficients are used as feature to the DNN that was substituted with the model in (58). These results about speaker classification are justified in the context of our study focused on low complexity

Table 2. LR performance (scenario 1) against two SNR (SNR1=2:1 and SNR2=1:1) and two batch size (BS1=32 and BS2=64) levels for the audio signal with 4 channels

Metrics –	**SNR 1** $BS1$	$BS2$	**SNR 2** $BS1$	$BS2$
Mean error (**azimuth**)	5.89 (v) 5.78 (t)	5.82 (v) 5.90 (t)	6.21 (v) 6.18 (t)	6.26 (v) 6.22 (t)
Median error (**azimuth**)	5.73 (v) 5.45 (t)	5.66 (v) 5.80 (t)	6.16 (v) 6.07 (t)	6.06 (v) 6.13 (t)
RMSE error (**azimuth**)	6.96 (v) 6.85 (t)	6.76 (v) 6.87 (t)	7.17 (v) 7.18 (t)	7.38 (v) 7.28 (t)
Accuracy (**class**)	0.89 (v) 0.88 (t)	0.88 (v) 0.88 (t)	0.83 (v) 0.83 (t)	0.81 (v) 0.81 (t)

validation (v) - *testing* (t)

Table 3. SRP-PHAT performance (scenario 1) against two SNR (SNR1=2:1 and SNR2=1:1) and two grid size (GS1=100 and GS2=200) values for the audio signal with 4 channels

Metrics –	**SNR 1** $GS1$	$GS2$	**SNR 2** $GS1$	$GS2$
Mean error (**azimuth**)	21.14	21.10	21.13	21.11
Median error (**azimuth**)	21.02	20.98	21.00	21.01
RMSE error (**azimuth**)	22.32	22.28	22.31	22.29

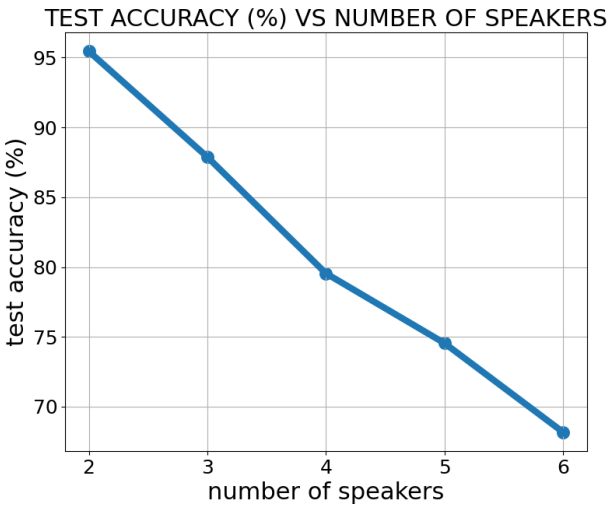

Figure 13. Classifier performance on the extended dataset (scenario 1) against the number of voices (2-6) with BS=32 and SNR=1:1 for the audio signal with 4 channels

methods where the acoustic data is heavily deteriorated by the UAV ego-noise. In fact, in this situation, the diversity in the distinctive characteristics of each voice is reduced.

For an analysis of the computational complexity, the single-frame computation time was measured in both the binary dataset (localization and recognition) and the multi-class dataset only recognition) for: 1) training of the DNN model; 2) testing of the DNN model; 3) SRP-PHAT method. The measured times in msec are in Table 4.

Table 4. Single-frame computation time (msec) for scenario 1 at two different batch size (BS1=32 and BS2=64) values for the DNN and grid size (GS1=100 and GS2=200) values for the SRP-PHAT, for both the binary and the multi-class dataset

Process	**BS/GS 1**	**BS/GS 2**
Binary dataset		
Training (DOA/CLASS)	4.94/5.56	4.64/4.98
Optimizer (DOA+CLASS)	6.94	6.63
Testing (DOA/CLASS)	3.78/3.95	3.28/3.44
SRP-PHAT (DOA)	26.67	30.69
Multi-class dataset		
Training (CLASS)	6.86	6.64
Optimizer (CLASS)	8.5	8.28
Testing (CLASS)	4.64	4.59
SRP-PHAT (DOA)	27.28	31.35

Table 5. LR performance (scenario 2) against two SNR (SNR1=2:1 and SNR2=1:1) and two batch size (BS1=32 and BS2=64) levels for the audio signal with 19 channels

Metrics –	**SNR 1** $BS1$	$BS2$	**SNR 2** $BS1$	$BS2$
Mean error (**azimuth**)	3.24 (v) 3.17 (t)	4.99 (v) 5.15 (t)	3.10 (v) 3.05 (t)	6.08 (v) 5.49 (t)
Median error (**azimuth**)	1.75 (v) 1.68 (t)	1.34 (v) 1.23 (t)	1.03 (v) 0.91 (t)	3.81 (v) 3.53 (t)
RMSE error (**azimuth**)	6.99 (v) 6.98 (t)	11.65 (v) 11.74 (t)	7.26 (v) 6.94 (t)	11.35 (v) 10.61 (t)
Accuracy (**class**)	0.93 (v) 0.89 (t)	0.96 (v) 0.94 (t)	0.94 (v) 0.90 (t)	0.96 (v) 0.87 (t)

validation (v) - *testing* (t)

Table 6. SRP-PHAT performance (scenario 2) against two SNR (SNR1=2:1 and SNR2=1:1) and two grid size (GS1=100 and GS2=200) values for the audio signal with 19 channels

Metrics –	**SNR 1** $GS1$	$GS2$	**SNR 2** $GS1$	$GS2$
Mean error (**azimuth**)	4.08	4.13	4.66	4.71
Median error (**azimuth**)	3.61	1.79	3.60	3.59
RMSE error (**azimuth**)	5.76	5.97	7.11	7.66

Scenario 2: (UAV LR with 19-channel signals) When trained with a batch size of 32, the system consistently produced low mean azimuth errors around 3.05° to 3.24°, and median errors below 1.75°, independent of SNR level. In contrast, using a larger batch size of 64 increased the azimuth mean error, particularly under higher noise, reaching up to 6.08°. The RMSE values supported this trend, ranging from about 6.9° to 11.7° depending on configuration. Classification results for distinguishing between different UAVs remained robust, with accuracies between 87% and 96%.

Table 7. Classifier performance on the extended dataset (scenario 2) against two SNR (SNR1=2:1 and SNR2=1:1) and two batch size (BS1=32 and BS2=64) levels for the audio signal with 19 channels

Metrics	**SNR 1**		**SNR 2**	
–	$BS1$	$BS2$	$BS1$	$BS2$
Accuracy	0.98 (v)	0.98 (v)	0.95 (v)	0.92 (v)
(**class**)	0.98 (t)	0.97 (t)	0.93 (t)	0.90 (t)

validation (v) - *testing* (t)

Table 8. Single-frame computation time (msec) for scenario 2 at two different batch size (BS1=32 and BS2=64) values for the DNN and grid size (GS1=100 and GS2=200) values for the SRP-PHAT, for both the binary and the multi-class dataset

Process	**BS/GS 1**	**BS/GS 2**
Binary dataset		
Training (DOA/CLASS)	1.66/1.00	2.18/1.18
Optimizer (DOA+CLASS)	3.72	5.19
Testing (DOA/CLASS)	1.50/0.92	1.28/1.00
SRP-PHAT (DOA)	558.55	1088.08
Multi-class dataset		
Training (CLASS)	0.96	1.12
Optimizer (CLASS)	4.08	4.32
Testing (CLASS)	0.96	1.02
SRP-PHAT (DOA)	566.05	1097.86

These results validate the system's ability to effectively estimate direction and recognize source types, even in realistic, noisy UAV conditions. Notably, a smaller batch size appears more favorable for localization performance in both scenarios.

The conventional SRP-PHAT method for acoustic target localization was applied to the UAV sound-based dataset in Scenario 2 to compare the performance with our DNN-based method. The time domain variant of the SRP-PHAT was considered with 200 grid points. The results are organized in Table 6 where two two ego-noise levels are considered.

The classifier performance when applied to the increased diversity dataset with 6 different UAVs can be seen from the resulting values of the metrics in Table 7 obtained by setting the batch size to 32 and 64 and the ego-noise ratio to two different levels.

For an analysis of the computational complexity, the single-frame computation time was measured in both the binary dataset (localization and recognition) and the multi-class dataset only recognition) for: 1) training of the DNN model; 2) testing of the DNN model; 3) SRP-PHAT method. The measured times in msec are in Table 8.

Visualizations of internal network representations (e.g., Fig. 14) further illustrate how the deep learning model processes features through its layers and consolidates information for final prediction, offering insight into the interpretability and functioning of the learned models.

Additionally, the inner representations of the most important of the network layers relative to the validation process of the Scenario 2, with batch size 64 and SNR 2:1, are also extracted and shown in Fig. 14. These images represent the way the signals are encoded throughout the layers in the DNN. Considering the DNN model in Fig. 4, the image in Fig. 14(a) is relative to the linear layer of the spectrum feature branch and the image in Fig. 14(b) is relative to the linear layer of the covariance matrix feature branch. The fusion layer generated the representation in Fig. 14(c). After the fusion block in the DNN model, a linear layer produces the sound DOA estimation at one head of the network, and another linear layer produces the sound class prediction at the other head of the network. Just before the last mentioned linear layer of the recognition head, the representation is generated as in Fig. 14(d). Specifically, the output of the DNN, localization and recognition predictions, is related to the encoded versions of the signals that travel through the DNN layers and can be visualized through these inner representations. Once the DNN is trained, its model parameters are frozen during the validation and testing steps and, then, the inner representations only change on variations of the input signal while the network parameters have no influence during these steps. This helps for both testing the DNN and investigating the way its output is formed.

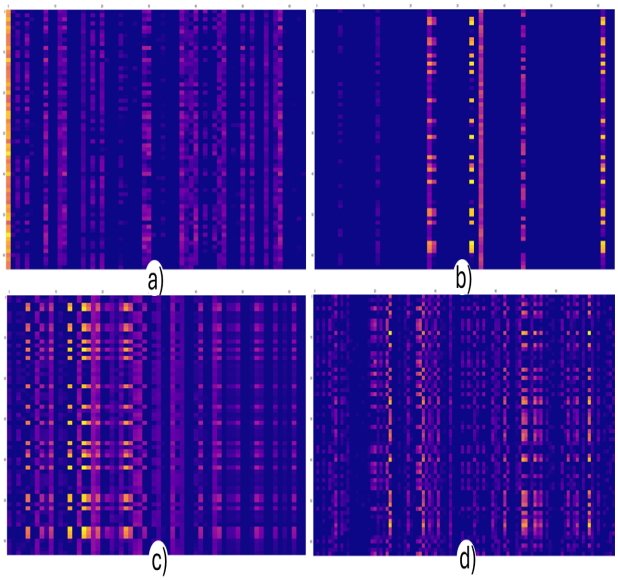

Figure 14. Inner representations of the most important of network layers for Scenario 2 (validation, batch size 64, SNR 2:1): a) spectrum feature linear layer b) covariance matrix feature linear layer c) fusion layer d) first of the two linear layers in the recognition head

7 Discussion

This work comprises three core components: system setup, dataset acquisition and processing, and deep learning-based acoustic scene interpretation for localization and recognition tasks. The proposed models are designed for real-world deployment in scenarios where traditional sensing may be limited, such as search-and-rescue or UAV monitoring. By analyzing multichannel audio data, the models extract spatial and semantic information from acoustic scenes, demonstrating robustness even under high ego-noise conditions. In fact, the network ability to detect

and discrimate the acoustic signature of the signals by means of the CNN and the other layers of the models provides the DNN the ability to mitigate the impact of noise.

The results are obtained under different signal-to-noise conditions (SNR1 and SNR2) and batch sizes (BS1 and BS2) for the two scenarios. It resulted that validation and testing produced similar values for each of the metrics, SNR, and BS, for both azimuth angle estimation and class prediction, and both scenarios. Overall, a higher ego-noise level (SNR2) slightly increases the azimuth prediction errors, although it is clearly more remarkable in Scenario 1 and with less influence in Scenario 2, and the same is observed on the classification accuracy. In Scenario 1, the batch size parameter has substantially no real influence. However, considering the testing (t) of the deep models in this scenario, a smaller batch size (BS1) results in slightly better performance for both localization and recognition. A higher batch size value (BS2) substantially worsens the localization performance in Scenario 2. It is observed that it has no real influence on the accuracy of the recognition in the same scenario. Then, having a smaller batch size (BS1) results in better performance, especially for localization. The conventional SRP-PHAT method for DOA estimation is used in the comparative analysis and the computational complexity is then evaluated by measuring the single-frame time for both our and the conventional methods to demonstrate the effective real-time feature of the DNN-based algorithm with respect to the conventional methods like the SRP-PHAT. A study over the dataset diversity is also conducted to show the classification performance in contexts with 2 to 6 speakers in Scenario 1 and with 6 different UAVs in Scenario 2.

Representations of the acoustic data as it propagates through the DNN are visualized for Scenario 2 to investigate the DNN and the way it produces the output predictions of the DOA and sound class. Understanding these representations results in a twofold advantage: bringing to light the inner processes at the layer level inside the DNN and optimizing the DNN based on these representations. Once fully investigated in the field of DL, this study may yield important findings and applications to be used in contexts like the acoustic scene awareness from UAV for real-time people search-and-rescue and counter flying object intrusion.

8 Conclusion and future work

This research presented experimental UAV-based acoustic scene awareness through LR in counter-UAV and search-and-rescue applications. By using acoustic sensors, a 19-microphone spherical array and a 4-microphone linear array, information from the acoustic scene can be extracted and analyzed. The motivation lies in the fact that analysis of the only acoustic scene through multi-channel audio processing for LR tasks still needs to be further investigated, especially in the context of the rapid developing DL. In particular, this study addresses the LR problem in scenarios where video information is not reliable or not available. In such cases, the audio-based algorithms should guarantee accurate results. A real experimental framework based on acoustic data acquisition by UAVs was utilized for LR of acoustic sources according to the two scenarios: 1) LR of target human voice from UAVs for search-and-rescue applications (the human voice is the source) and 2) LR of target UAVs from UAVs for counter-UAV applications (the propeller noise is the source). In a data-driven framework, two feature-based DNN models based on CNN networks were investigated and utilized to analyze the data gathered from the corresponding acoustic scene and perform the LR tasks about DOA estimation (localization) and class prediction (recognition). These models are capable of processing the multi-channel acoustic signals even in the presence of UAV ego-noise.

The results demonstrated the validity of the algorithm for both scenarios at different signal-to-noise levels. Comparison with the conventional SRP-PHAT method for DOA estimation, computational time analysis to demonstrate the real-time feature of the DNN-based approach, and multi-class classification based on an increased diversity dataset with 2 to 6 speakers in Scenario 1 and 6 UAVs in Scenario 2 are also conducted in the LR study. Future work will expand the range of acoustic sources and sensing devices to improve model generalizability. Additionally, ongoing research will explore both the optimization of deep neural network architectures, hyperparameters, and interpretability through internal feature visualization and the potential of combining multi-channel audio processing with deep learning. The environmental influence caused by reverberation from walls, buildings, trees, and other obstacles will also be investigated in future work to fully characterize its impact on the LR task. To improve the suitability of the LR method for UAV embedded platforms, strategies for reducing the computational complexity such as model pruning, quantization, and signal downsampling can be considered. These efforts aim to enhance the practical deployment of audio-based scene awareness in critical applications such as search-and-rescue and UAV intrusion detection.

Acknowledgements

Partial financial support was received from PNRR DD 3277 del 30 dicembre 2021 (PNRR Missione 4, Componente 2, Investimento 1.5) - iNEST.

Dr. Andrea Gulli from University of Udine took part in the "human voice generated by speech" data collection and dataset preparation used in this work.

Statements and declarations

- **Author contributions:** All authors have had an active part in the study and the manuscript preparation. All authors have approved the manuscript, and agree with its submission to the Integrated Computer-Aided Engineering journal.

- **Ethical considerations:** All the research meets the ethical guidelines and legal requirements specified by the Integrated Computer-Aided Engineering journal.

- **Consent to participate:** Not applicable.

- **Consent for publication:** Not applicable.

- **Funding:** The authors received no financial support for the research, authorship, and/or publication of this article.

- **Declaration of conflicting interest:** The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

[1] Zhang J and Zhang Y. A Method for UAV Reconnaissance and Surveillance in Complex Environments. In *2020 6th International Conference on Control, Automation and Robotics (ICCAR)*. pp. 482–485. DOI:10.1109/ICCAR49639.2020.9107972.

[2] Huang H and Savkin AV. Reactive 3D deployment of a flying robotic network for surveillance of mobile targets. *Computer Networks* 2019; 161: 172–182. DOI:https://doi.org/10.1016/j.comnet.2019.06.020.

[3] Cecchinato N, Scagnetto I, Toma A et al. A broadcast sub-GHz framework for unmanned aerial vehicles clock synchronization. *Integrated Computer-Aided Engineering* 2024; 31(1): 59–75. DOI:10.3233/ICA-230723.

[4] Nakadai K, Kumon M, Okuno HG et al. Development of microphone-array-embedded UAV for search and rescue task. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. pp. 5985–5990. DOI:10.1109/IROS.2017.8206494.

[5] Cecchinato N, Toma A, Scagnetto I et al. An Integrated Monitoring System for Aerial Drones and Underwater ROVs. In *2023 IEEE International Workshop on Technologies for Defense and Security (TechDefense)*. pp. 187–191. DOI:10.1109/TechDefense59795.2023.10380896.

[6] Toma A, Cecchinato N, Ferrin G et al. MAV-Link-Based Control and Coordination of a Multi-Drone Cluster for Intelligence, Surveillance and Reconnaissance Tasks. *Unmanned Systems* 2024; 13(04): 1027–1040. DOI:10.1142/S2301385025500633.

[7] Toma A, Cecchinato N, Drioli C et al. Onboard Audio and Video Processing for Secure Detection, Localization, and Tracking in Counter-UAV Applications. *Procedia Computer Science* 2022; 205: 20–27. DOI:https://doi.org/10.1016/j.procs.2022.09.003. 2022 International Conference on Military Communication and Information Systems (ICMCIS).

[8] Liu H, Wei Z, Chen Y et al. Drone Detection Based on an Audio-Assisted Camera Array. In *2017 IEEE Third International Conference on Multimedia Big Data (BigMM)*. pp. 402–406. DOI:10.1109/BigMM.2017.57.

[9] Yamazaki Y, Premachandra C and Perea CJ. Audio-Processing-Based Human Detection at Disaster Sites With Unmanned Aerial Vehicle. *IEEE Access* 2020; 8: 101398–101405. DOI:10.1109/ACCESS.2020.2998776.

[10] Zhang B, Masahide K and Lim H. Sound Source Localization and Interaction based Human Searching Robot under Disaster Environment. In *2019 SICE International Symposium on Control Systems (SICE ISCS)*. pp. 16–20. DOI:10.23919/SICEISCS.2019.8758766.

[11] Rosati R, Fabiani M, Pierdicca R et al. An automated workflow based on UAV imagery and deep learning methods for monitoring excavation area work. *Integrated Computer-Aided Engineering* 2025; DOI:10.1177/10692509251340464.

[12] Gąsienica-Józkowy J, Knapik M and Cyganek B. An ensemble deep learning method with optimized weights for drone-based water rescue and surveillance. *Integrated Computer-Aided Engineering* 2021; 28(3): 221–235. DOI:10.3233/ICA-210649.

[13] Jiang T, Frøseth GT, Rønnquist A et al. A visual inspection and diagnosis system for bridge rivets based on a convolutional neural network. *Computer-Aided Civil and Infrastructure Engineering* 2024; 39(24): 3786–3804. DOI:https://doi.org/10.1111/mice.13274.

[14] Martins GB, Papa JP and Adeli H. Deep learning techniques for recommender systems based on collaborative filtering. *Expert Systems* 2020; 37(6): e12647. DOI:https://doi.org/10.1111/exsy.12647.

[15] Hassanpour A, Moradikia M, Adeli H et al. A novel end-to-end deep learning scheme for classifying multi-class motor imagery electroencephalography signals. *Expert Systems* 2019; 36(6): e12494. DOI:https://doi.org/10.1111/exsy.12494.

[16] Kolamunna H, Dahanayaka T, Li J et al. DronePrint: Acoustic Signatures for Open-set Drone Detection and Identification with Online Data. *Association for Computing Machinery* 2021; 5(1). DOI:10.1145/3448115.

[17] Al-Emadi S, Al-Ali A and Al-Ali A. Audio-Based Drone Detection and Identification Using Deep Learning Techniques with Dataset Enhancement through Generative Adversarial Networks. *Sensors* 2021; 21(15). DOI:10.3390/s21154953.

[18] Kolamunna H, Li J, Dahanayaka T et al. AcousticPrint: acoustic signature based open set drone identification. In *Proceedings of the 13th ACM Conference on Security and Privacy in Wireless and Mobile Networks*. WiSec '20, New York, NY, USA: Association for Computing Machinery. ISBN 9781450380065, p. 349–350. DOI:10.1145/3395351.3401700.

[19] Choi J and Chang J. Convolutional Neural Network-based Direction-of-Arrival Estimation using Stereo Microphones for Drone. In *2020 International Conference on Electronics, Information, and Communication (ICEIC)*. pp. 1–5. DOI:10.1109/ICEIC49074.2020.9051364.

[20] He W, Motlicek P and Odobez JM. Neural Network Adaptation and Data Augmentation for Multi-Speaker Direction-of-Arrival Estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 2021; : 1–1DOI:10.1109/TASLP.2021.3060257.

[21] Salvati D, Drioli C and Foresti GL. Exploiting CNNs for Improving Acoustic Source Localization in Noisy and Reverberant Conditions. *IEEE Transactions on Emerging Topics in Computational Intelligence* 2018; 2(2): 103–116. DOI:10.1109/TETCI.2017.2775237.

[22] Wang L and Cavallaro A. Time-Frequency Processing for Sound Source Localization from a Micro Aerial Vehicle. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 496–500. DOI:10.1109/ICASSP.2017.7952205.

[23] Farhadi S, Corrado M and Ventura G. Automated acoustic event-based monitoring of prestressing tendons breakage in concrete bridges. *Computer-Aided Civil and Infrastructure Engineering* 2024; 39(24): 3700–3720. DOI:https://doi.org/10.1111/mice.13321.

[24] Varanasi V, Gupta H and Hegde RM. A Deep Learning Framework for Robust DOA Estimation Using Spherical Harmonic Decomposition. *IEEE/ACM Transactions on*

Audio, Speech, and Language Processing 2020; 28: 1248–1259. DOI:10.1109/TASLP.2020.2984852.

[25] Chakrabarty S and Habets EAP. Multi-Speaker DOA Estimation Using Deep Convolutional Networks Trained With Noise Signals. *IEEE Journal of Selected Topics in Signal Processing* 2019; 13(1): 8–21. DOI:10.1109/JSTSP.2019.2901664.

[26] Wang ZQ, Zhang X and Wang D. Robust Speaker Localization Guided by Deep Learning-Based Time-Frequency Masking. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 2019; 27(1): 178–188. DOI:10.1109/TASLP.2018.2876169.

[27] Cobos M, Antonacci F, Alexandridis A et al. A Survey of Sound Source Localization Methods in Wireless Acoustic Sensor Networks. *Wirel2017ess Communications and Mobile Computing* 2017; 2017(1): 3956282. DOI:https://doi.org/10.1155/2017/3956282.

[28] Oh S, Go YJ, Lee J et al. Sound source positioning using microphone array installed on a flying drone. *The Journal of the Acoustical Society of America* 2016; 140: 3422–3422. DOI:10.1121/1.4971007.

[29] Salvati D, Drioli C, Ferrin G et al. Beamforming-Based Acoustic Source Localization and Enhancement for Multirotor UAVs. In *2018 26th European Signal Processing Conference (EUSIPCO)*. pp. 987–991. DOI:10.23919/EUSIPCO.2018.8553514.

[30] Yamazaki Y, Tamaki M, Premachandra C et al. Victim Detection Using UAV with On-board Voice Recognition System. In *2019 Third IEEE International Conference on Robotic Computing (IRC)*. pp. 555–559. DOI:10.1109/IRC.2019.00114.

[31] Abdulghani MM, Walters WL and Abed KH. Voice Signature Recognition for UAV Pilots Identity Verification. In *2023 International Conference on Computational Science and Computational Intelligence (CSCI)*. pp. 125–129. DOI:10.1109/CSCI62032.2023.00026.

[32] Morito T, Sugiyama O, Kojima R et al. Partially Shared Deep Neural Network in sound source separation and identification using a UAV-embedded microphone array. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. pp. 1299–1304. DOI:10.1109/IROS.2016.7759215.

[33] Banerjee A, Nilhani A, Dhabal S et al. Chapter 15 - A novel sound source localization method using a global-best guided cuckoo search algorithm for drone-based search and rescue operations. In Koubaa A and Azar AT (eds.) *Unmanned Aerial Systems*. Advances in Nonlinear Dynamics and Chaos (ANDC), Academic Press. ISBN 978-0-12-820276-0, 2021. pp. 375–415. DOI:https://doi.org/10.1016/B978-0-12-820276-0.00022-4.

[34] Strauss M, Mordel P, Miguet V et al. DREGON: Dataset and Methods for UAV-Embedded Sound Source Localization. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. pp. 1–8. DOI:10.1109/IROS.2018.8593581.

[35] Deleforge A, Di Carlo D, Strauss M et al. Audio-Based Search and Rescue With a Drone: Highlights From the IEEE Signal Processing Cup 2019 Student Competition [SP Competitions]. *IEEE Signal Processing Magazine* 2019; 36(5): 138–144. DOI:10.1109/MSP.2019.2924687.

[36] Solis ER, Shashev DV and Shidlovskiy SV. Implementation of Audio Recognition System for Unmanned Aerial Vehicles. In *2021 International Siberian Conference on Control and Communications (SIBCON)*. pp. 1–8. DOI:10.1109/SIBCON50419.2021.9438906.

[37] McCoy J, Rawal A, Rawat DB et al. Ensemble Deep Learning for Sustainable Multimodal UAV Classification. *IEEE Transactions on Intelligent Transportation Systems* 2023; 24(12): 15425–15434. DOI:10.1109/TITS.2022.3170643.

[38] Lee H, Han S, Byeon JI et al. CNN-Based UAV Detection and Classification Using Sensor Fusion. *IEEE Access* 2023; 11: 68791–68808. DOI:10.1109/ACCESS.2023.3293124.

[39] Sedunov A, Haddad D, Salloum H et al. Stevens Drone Detection Acoustic System and Experiments in Acoustics UAV Tracking. In *2019 IEEE International Symposium on Technologies for Homeland Security (HST)*. pp. 1–7. DOI:10.1109/HST47167.2019.9032916.

[40] Hu X, Yang M, Liu C et al. Intelligent Unmanned Defense System for Autonomous Interception of UAVs Based on Improved Acoustic Source Localization Algorithm. *IEEE Access* 2025; : 1–1DOI:https://doi.org/10.1109/ACCESS.2025.3575959.

[41] Chang X, Yang C, Shi X et al. Feature Extracted DOA Estimation Algorithm Using Acoustic Array for Drone Surveillance. In *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*. pp. 1–5. DOI:10.1109/VTCSpring.2018.8417601.

[42] Chevtchenko SF, Rodríguez BJ, Vale R et al. Drone-Based Sound Source Localization: A Systematic Literature Review. *IEEE Access* 2025; 13: 94256–94274. DOI:10.1109/ACCESS.2025.3572478.

[43] Martinez-Carranza J and Rascon C. A Review on Auditory Perception for Unmanned Aerial Vehicles. *Sensors* 2020; 20(24). DOI:10.3390/s20247276.

[44] Salvati D, Drioli C and Foresti GL. Iterative Diagonal Unloading Beamforming for Multiple Acoustic Sources Localization Using Compact Sensor Arrays. *IEEE Sensors Journal* 2021; 21(13): 15080–15089. DOI:10.1109/JSEN.2021.3074622.

[45] Insausti X, Hogstad BO and Pätzold M. Modelling and Simulation of Ego-Noise of Unmanned Aerial Vehicles. In *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*. pp. 1–5. DOI:10.1109/VTC2020-Spring48590.2020.9128572.

[46] Moshkov P. Study of the Propellers Noise of Light Aircraft under Static Conditions. In *2022 International Conference on Dynamics and Vibroacoustics of Machines (DVM)*. pp. 1–6. DOI:10.1109/DVM55487.2022.9930902.

[47] Podsędkowski M, Konopiński R and Lipian M. Sound noise properties of variable pitch propeller for small UAV. In *2022 International Conference on Unmanned Aircraft Systems (ICUAS)*. pp. 1025–1029. DOI:10.1109/ICUAS54217.2022.9836100.

[48] Kingan MJ, McKay RS, Wu Y et al. Unmanned Aerial Vehicle Propeller Noise. In Doolan C, Moreau D and Wills A (eds.) *Flinovia—Flow Induced Noise and Vibration Issues and Aspects—IV*. Cham: Springer Nature Switzerland, 2025. ISBN 978-3-031-73935-4, pp. 103–118.

[49] Toma A, Salvati D, Drioli C et al. Efficient Detection and Localization of Acoustic Sources with a low complexity CNN

network and the Diagonal Unloading Beamforming. In *2022 International Joint Conference on Neural Networks (IJCNN)*. pp. 1–8. DOI:10.1109/IJCNN55064.2022.9892709.

[50] Toma A, Salvati D, Drioli C et al. CNN-Based Processing of Acoustic and Radio Frequency Signals for Speaker Localization from MAVs. In *Proc. Interspeech*. pp. 2147–2151. DOI:10.21437/Interspeech.2021-886.

[51] Toma A, Cecchinato N, Drioli C et al. CNN-based processing of radio frequency signals for augmenting acoustic source localization and enhancement in UAV security applications. In *2021 International Conference on Military Communication and Information Systems (ICMCIS)*. pp. 1–5. DOI:10.1109/ICMCIS52405.2021.9486424.

[52] Parrot Bebop Drone - User guide. https://www.parrot.com/assets/s3fs-public/2021-09/bebop-drone_user-guide_uk_v.3.4.pdf. Accessed: 2025-06-20.

[53] DJI Matrice 200 - User Manual. https://dl.djicdn.com/downloads/M200/20201120/M200_User_Manual_EN_20201120.pdf. Accessed: 2025-06-20.

[54] DJI Flame Wheel ARF KIT. https://www-v1.dji.com/flame-wheel-arf.html. Accessed: 2025-06-20.

[55] PlayStation Eye - PS3 Eye. https://en.wikipedia.org/wiki/PlayStation_Eye. Accessed: 2025-06-20.

[56] ZYLIA ZM-1 3rd order ambisonics microphone array. https://www.zylia.co/zylia-zm-1-microphone.html. Accessed: 2025-06-20.

[57] Toma A. Acoustic UAV Datasets for Deep Acoustic Learning on Unmanned Aerial Vehicles for Real-Time Human and Drone Detection. https://lambda-iot.uniud.it/UAV_DeepAcousticLocalizationRecognition_Dataset, 2025.

[58] Speaker Identification AI. https://github.com/ManuOtel/Speaker-Identification-AI/tree/main. Accessed: 2025-10-14.